١٥

تكييف الاختبارات التربوية والنفسية للتقييم عبر الثقافات

منتدى سور الأزبكية
www.books4all.net

تحرير
رونالد ك. هامبلتون
بيتر ف. ميريندا
تشارلز د. سبيلبيرغر

مكتبات ونشر العبيكان
Obeikan
Publishers & Booksellers

وزارة التعليم العالي
المملكة العربية السعودية

منتدى سور الأزبكية

WWW.BOOKS4ALL.NET

https://twitter.com/SourAlAzbakya

https://www.facebook.com/books4all.net

منتدى سور الأزبكية
www.books4all.net

تكييف الاختبارات التربوية
والنفسية للتقييم عبر الثقافات

تكييف الاختبارات التربوية والنفسية للتقييم عبر الثقافات

منتدى سور الأزبكية
www.books4all.net

تحرير

رونالد ك. هامبلتون

جامعة ماستشوست/ أمهرس

بيتر ف. ميرينـــــدا

جامعة رود آيلاند

تشارلز د. سبيلبيرغر

جامعة جنوب فلوريدا

نقله إلى العربية

هالة برمدا

راجعه

د. مصطفى عشوي

مكتبة العبيكان
Obeikan
Publishers & Booksellers

Original Title:

Adapting Educational and Psychological Tests For Cross-Cultural Assessment

by:

Ronald K. Hambleton, Peter F. Merenda, Charles D. Spielberger

Copyright © 2005 by Lawrene Erlbaum Associates, Inc.

ISBN 0 - 8058 - 3025 - 1

All rights reserved. Authorized translation from the English language edition
published by: Lawrence Erlbaum Associates,

حقوق الطبعة العربية محفوظة للعبيكان بالتعاقد مع لورانس إيرلبوم أسوسيتس، ناشرون

© مكتبة العبيكان 1427هـ ـ 2006م

المملكة العربية السعودية، شمال طريق الملك فهد مع تقاطع العروبة، ص.ب. 62807 الرياض 11595

Obeikan Publishers, North King Fahd Road, P.O. Box 62807, Riyadh 11595, Saudi Arabia

الطبعة العربية الأولى 1427هـ ـ 2006م

ISBN 4 - 888 - 40 - 9960

ⓒ مكتبة العبيكان، 1426هـ

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

هامبلتون، رونالد ك

تكييف الاختبارات النفسية للتقويم عبر الثقافات. / رونالد ك؛ هامبلتون؛ بيتر اف ميرندا؛ تشارلز دسبيلبير غر؛ هالة برمدا. - الرياض 1426هـ

494 ص؛ 16.5× 24 سم

ردمك: 4 - 888 - 40 - 9960

١ ـ الاختبارات والمقاييس التربوية — 2 – الاختبارات النفسية

أ. ميرندا، بيتر اف (مؤلف مشارك) — ب. سبيلبير غر، تشارلز د (مؤلف مشارك)

ج. برمدا، هالة (مترجم) — د. العنوان

ديوي: 371.27 — 1426 / 7721

رقم الإيداع: 1426 / 7721

ردمك: 4 - 888 - 40 - 9960

جميع الحقوق محفوظة. ولا يسمح بإعادة إصدار هذا الكتاب أو نقله في أي شكل أو واسطة، سواء أكانت إلكترونية أو ميكانيكية، بما في ذلك التصوير بالنسخ «فوتوكوبي»، أو التسجيل، أو التخزين والاسترجاع، دون إذن خطي من الناشر.

All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

بسم الله الرحمن الرحيم

تقديم معالي وزير التعليم العالي

الحمد لله والصلاة والسلام على رسول الله وبعد:

تحرص وزارة التعليم العالي في المملكة العربية السعودية على تشييد بنية متينة للتعليم العالي في المملكة تأخذ في الحسبان متطلبات مجتمعها وثقافته الإسلامية العريقة، وفي الوقت نفسه تحاكي أنظمة التعليم العالي العالمية. وكان الغرض الأساس للسعي وراء هذا الهدف هو تطوير العملية التعليمية، وكذلك تطوير النظام الإداري المصاحب خاصة في ضوء الطفرة المعلوماتية والعولمة والمنافسة الشديدة بين مؤسسات التعليم العالي على المستويات المحلية والإقليمية والدولية.

ونظراً لما حققه التعليم العالي في المملكة العربية السعودية من تطور كمّي ونوعي بدعم سخي من حكومتنا الرشيدة بقيادة خادم الحرمين الشريفين الملك عبدالله بن عبدالعزيز، وسمو ولي عهده الأمين، الأمير سلطان بن عبدالعزيز - يحفظهما الله- فقد ظهرت الحاجة بشكل أكبر لتوفير المصادر المختلفة لتعزيز توعية الأفراد العاملين في حقل التعليم العالي بما ينشر في هذا المجال باللغات الأجنبية. لذا، رأت وزارة التعليم العالي ترجمة عدد من الكتب ذات العلاقة بمجالات التطوير الأكاديمي وتقديمها باللغة العربية لتكون في متناول جميع العاملين في القطاع الأكاديمي. ونظرا لقلة مثل هذه الكتب في المكتبة العربية، فقد سعت الوزارة إلى توفيرها بشكل سريع وفاعل، وعليه كان مشروع الترجمة هذا. ولقد قامت الوزارة باختيار كتب تحوي دراسات حازت قبولا وانتشارا في الكثير من المؤسسات التعليمية ذات الشهرة العالمية وأنجزت بأيدي عدد من الأكاديميين والإداريين المهتمين بالتطوير في التعليم العالي. وعالجت الدراسات في هذه الكتب قضايا متعلقة بكل من تطوير مهارات الأساتذة ورؤساء الوحدات الأكاديمية والإداريين في

أكثر الجامعات العالمية تقدما. كما تناولت هذه الكتب قضايا مثل: التعليم الإلكتروني، والتعليم عن بعد، ومهارات التعليم والتعلم، وتقنيات التعليم الحديثة، والتخطيط الاستراتيجي الخاص بالتعليم، والاختبارات والتقويم، ومواءمة مخرجات التعليم العالي لسوق العمل، وتحقيق الجودة في مدخلات ومخرجات التعليم العالي وغير ذلك من الموضوعات ذات العلاقة.

ووقع اختيار الوزارة على مكتبة العبيكان للنشر للتعاون معها في نشر ترجمات هذه السلسلة من الكتب الأكاديمية المتخصصة وذلك لما لهذه المكتبة من خبرة وتميز في مجال النشر وفي ميداني التأليف والترجمة والكفاءة في الأداء. وقامت مكتبة العبيكان بمهمة الاتفاق مع الناشرين للكتب الأجنبية ومن ثم ترجمتها وتقديمها للقارئ بالشكل المناسب، وقد تم مراجعة هذه الكتب من قبل فرق أكاديمية متخصصة.

وتأمل الوزارة بأن تكون بهذا المشروع قد أسهمت بوضع دليل متكامل من الدراسات المهمة والمشروعات والأفكار ذات العلاقة بتطوير التعليم العالي بين أيدي جميع أعضاء الهيكل الأكاديمي و الإداري في الجامعات ابتداء من مديري الجامعات إلى أول الصاعدين على سلم التعليم والإدارة فيها.

وإذ تقدم هذه الكتب وأفكارها خلاصة تجارب المجتمعات الأكاديمية المتطورة في هذا المجال فإنها لا تقلّل من الخبرات ولا التجارب الميدانية المحلية لدينا، وتلك المستمدة من ديننا الحنيف وثقافتنا بل إنها ستعزز دور المجتمع الأكاديمي والإسهام في بناء وطننا الكريم، كما ستساعدنا على التخلص من الأخطاء التي مررنا بها أو وقعت لغيرنا فنتجنب تكرارها.

ولا يفوتني أن أشكر معالي الدكتور خالد بن صالح السلطان مدير جامعة الملك فهد للبترول والمعادن، وسعادة الدكتور سهل بن نشأت عبدالجواد، عميد التطوير

الأكاديمي في الجامعة، وجميع من عمل معهم على جهودهم المباركة لإخراج هذا المشروع إلى أن أصبح واقعا ملموساً وجهداً متميزاً، والذي سيكون له -بإذن الله- مردود إيجابي على المجتمع.

وفي الختام يسرنا أن تنشر وزارة التعليم العالي في المملكة العربية السعودية بالتعاون مع مكتبة العبيكان للنشر هذه السلسلة من ترجمات الكتب الأكاديمية المتخصصة، ونأمل أن تكون دليلا معرفيا يسهم في التطوير والتنمية، وذلك بجانب ما توافر في السابق لننطلق للمستقبل بأحسن ما توافر لدينا من خبراتنا الخاصة وما نتعلمه من تجارب الآخرين في جوانب البحث العلمي والأكاديمي في العالم... والله ولي الموفق،،،

الدكتور/ خالد بن محمد العنقري
وزير التعليم العالي في المملكة العربية السعودية

الفهرس

الموضوع	الصفحة
مقدمة	13
• **القسم الأول: تكييف الاختبارات التربوية والنفسية عبر الثقافات: مواضيع نظرية ومنهجية**	19
1- موضوعات، وخطط، إرشادات تقنيه لتكييف الاختبارات للغات وثقافات متعددة.	
رونالد. ك. هامبلتون.	21
2- مسائل مفاهيمية ومنهجية في تكييف الاختبارات.	
فونس. ج. ردي فيجر ويب هـ. بورتينغا.	69
3- موضوعات أخلاقية منتقاة ذات علاقة بتكييف الاختبار.	
توماس أوكلاند.	103
4- طرق إحصائية لتحديد العيوب في عملية تكييف الاختبارات.	
ستيفن ج. سيرغي، ليان باتسولا، رونالد ك. هامبلتون.	141
5- استخدام ثنائي اللغة لتقييم التشابه بين صيغ لغوية مختلفة لاختبار ما.	
ستيفن ج، سيرسي.	173
6- إرساء قواعد لمقارنة الدرجات لاختبارات معطاة بلغات مختلفة.	
ليندا ك. كوك، واليسيا شميدت كاسكالار	205
7- تكييف اختبارات الإنجاز والأهلية: مراجعة موضوعات منهجية.	
ليندا ل. كوك، اليسيا شميدت كاسكالار، كاثرين براون.	245
• **القسم الثاني: تكييف التقاطع اللغوي للاختبارات النفسية والتربوية: تطبيقات على اختبارات تتعلق بالإنجاز والأهلية والشخصية.**	275

٨- تكييف الاختبار في برنامج مصدّق «معتمد» واسع النطاق.

سيندي ت. فيتزجيرالد. ــــــــ 277

٩- تحويل مقياس فكسلر لقياس ذكاء البالغين: محاولة تكييف اختبار مبكرة ذات نتائج قيمّة.

كارلوس ي. مالدونادو، كيرت ف. كسينجر. 301

١٠- تطوير الاختبارات للاستعمال في اللغات والثقافات المتعددة: التماس للتطوير المتزامن.

نوربرت ك. تانزر. 329

١١- القياس النفسي للتكييف: تقييم تعادل القياس عبر اللغات والثقافات

يتز دراسكو، تاهيرا برويسن. ــــــــ 363

١٢- إنشاء وتكييف وتقويم صحة اختبارات القبول بلغات متعددة: الحالة الإسرائيلية.

مايكل بيلر. 405

١٣- التكييف عبر الثقافات للاختبارات التربوية والنفسية.

بيتر ف. ميريندا. ــــــــ 431

١٤- تقييم التقاطع الثقافي للحالات الوجدانية والسمات الشخصية.

تشارلز د. سبيلبرغر، مانوليت س. موسكو سو، وتوماس م. برونر. ــــــــ 461

مقدمة

في عام 1989 قرأت بالصدفة تقريراً عن مقارنة تقدير إنجازات حسابية لطلاب مدارس من خمس دول. فوجئت بالنتائج وبدأت أتساءل عن إمكانية تأثير عوامل منهجية مختلفة على تلك النتائج: نوعية الطلاب المنتقاة في كل من الدول المشاركة، الاختبارات الدقيقة لمحتويات وتصميم الاختيار، لكن كان أكبر اهتمامي بالطريقة التي تتم فيها ترجمة الاختبارات من اللغة الانجليزية إلى اللغات الأخرى التي استخدم فيها الاختبار في تلك الدراسة. إن الدراسات العالمية للإنجازات التربوية مهمة جداً لصناع السياسة التربوية لكنها لن تكون كذلك إذا شوهت عوامل منهجية مصداقية النتائج. ما صدمني حقاً احتمال أن يكون أسباب تلك النتائج هو أن المترجمين جعلوا تلك الاختبارات من غير قصد أسهل أو أصعب. ماذا كانت مؤهلات هؤلاء المترجمين؟ كم أعطوا من الوقت لإتمام العمل؟ ما هو الدليل التجريبي الذي جمع لمساندة الاختبار الموازي في اللغات المتعددة.

اتصلت بوكالة الاختبارات المسؤولة عن إقامة تلك الدراسات لمناقشة طرق الترجمة. لسوء الحظ لم يكن انطباعي جيداً عن التفاصيل التي زودوني بها عن كيفية ترجمة تلك الاختبارات، كيف دققت الجوانب اللغوية والنفسية حسب المجموعات المتعددة اللغات والثقافات. حسب تدقيقي الشخصي الذي قمت به عن ممارسة الترجمة الجيدة للاختبارات أصبت بخيبة أمل للمستوى الضعيف نسبياً في معرفة المناهج مقارنة بمعرفة الاختبارات في حقول مهمة أخرى مثل كيفية تطوير الاختبارات، موازنة درجات الاختبار وإحداث معيار لدرجات الاختبار. كانت تلك المواجهة الأولى لي مع عالم الاختبارات عبر الثقافات المختلفة وأدركت أن هناك الكثير من الأشياء التي يجب القيام بها.

في عام 1991 أبديت قلقي حول منهجية ترجمة الاختبارات إلى اللجنة التابعة إلى الهيئة الدولية للاختبارات (ITC)، اليوم أصبحت تلك الهيئة مؤسسة للجمعيات النفسية الوطنية، لوكالات الاختبارات، ولأشخاص آخرين في ذلك المجال وتقوم الهيئة بالإشراف على تحسين ممارسة الاختبارات في العالم.

قررت تلك الهيئة (ITC) تشكيل هيئة عالمية من العلماء والخبراء في الاختبارات لوضع خطوط عريضة أساسية لمنهجية ترجمة وتكييف الاختبارات. لحسن الحظ استطعنا الحصول على بعض المساعدات المالية لتلك الهيئة من المركز القومي للاحصاءات التربوية ومن مجلس الجامعات في الولايات المتحدة. كما تمكنا من جذب اهتمام عدد من المؤسسات العالمية إلى عمل الهيئة فقامت تلك المؤسسات بتزويدها بعدد من الأعضاء العاملين.

من هذه المؤسسات: الجمعية الأوروبية للتقويم النفسي، المجموعة الأوروبية لناشري الاختبارات، الجمعية العالمية لعلم النفس عبر الثقافات، الجمعية العالمية لعلم النفس التطبيقي، الجمعية العالمية لتقويم الإنجازات التربوية، الجمعية العالمية للاختبارات اللغوية والاتحاد العالمي للعلوم النفسية.

قام أعضاء تلك الهيئة (ITC) بالعمل بجد طول ثلاث سنوات وعقد كثير من الاجتماعات لتنظيم التطورات التقنية التي تم القيام بها في مجال ترجمة وتعديل الاختبارات. أخيراً قدمت الهيئة التقرير الختامي الذي يوفر للعاملين في مجال الترجمة 22 خطاً من الخطوط الأساسية للعمل (دليل هيئة الاختبارات العالمية في تكييف الاختبارات) يوجد ذلك الدليل ضمن الفصل الأول.

في الوقت الذي كانت مسودة ذلك الدليل تعرض على اللجنة لوضع الملاحظات عليه، قررنا أنا وتوم أولاند من جامعة فلوريدا، الولايات المتحدة وعضو آخر في الهيئة تنظيم مؤتمر لتقديم ذلك الدليل. أقيم ذلك المؤتمر الذي رعته الهيئة في جامعة جورج تاون (الولايات المتحدة) في ربيع عام 1999 . كان الحضور في ذلك

المؤتمر عالياً جداً. ركز ذلك المؤتمر على الدليل لترجمة وتكييف الاختبارات. وكانت الهيئة قد توقعت أن ذلك الدليل سيلقى ترحيباً في حقل الاختبارات وسيكون إضافة مهمة في المطبوعات الحديثة.

في ذات الوقت الذي أقيم فيه المؤتمر أبدى البروفسور تشارلز سبيلبيرغ (عضو في الهيئة ومساهم في وضع الدليل) وبيتر ميريندا موافقتهم على المساعدة في التحضير لكتاب يلقي الضوء على التقدم التطبيقي المهم في حقل ترجمة وتكييف الاختبارات، وكان سبيليبرغ نفسه قد ساهم في 50 ترجمة مستخدماً وسيلته الخاصة (البيان التفصيلي في حالات القلق) وكان ميريندا عضواً ناشطاً في ترجمة الأبحاث خلال ممارسته مهنته. اجتمعنا نحن الثلاثة لإخراج ذلك الكتاب، وهو مجموعة من المحاضرات التي ألقيت في مؤتمر جورج تاون وفصول إضافية أخرى لتوفير تغطية أكثر وضوحاً للموضوع.

إن الفصل الأول الذي كتبته بنفسي يهدف إلى تقديم الخطوط الأساسية في تكييف الاختبارات بالإضافة إلى وصف بعض الأمور التي قد تظهر خلال عملية ترجمة وتكييف الاختبارات.

أعد الفصل الثاني "البروفسور فون فان دو فيفر" و"يب بورتينغا" من جامعة تيلبرغ في هولندا ويدور حول قضايا عن مفاهيم ومناهج في تكييف الاختبارات. لو لم يكن الهدف تقديم الخطوط الأساسية في الفصل الأول كان على الفصل الثاني أن يكون الأول لأن الكاتبين قدموا خطوطاً عريضة لفهم عملية الترجمة والتكييف المتعلقة في كل الكتاب. أعد البروفسور توم أوكلاند الفصل الثالث حيث يناقش كل القضايا المهمة عن الأخلاقيات وتكييف الاختبارات. وكان جوهر هذا البحث قلقه من مصداقية درجات الاختبار في البيئات المتعددة الثقافات.

تقدم الفصول 4-7 نماذج رائعة للتقدم في منهجية ترجمة وتكييف الاختبارات، يقدم الفصل الرابع، الذي كتبه ستيف سيرغي، لينا باستولا (التي تعمل حالياً في

الاختبارات التعليمية في الولايات المتحدة) وأنا، عرضاً شاملاً لطرق مطابقة أغلاط في بعض بنود الاختبار التي تحدث أثناء عملية ترجمة وتكييف الاختبار. في الفصل الخامس كان البروفسور سيرغي مهيأ جداً لمناقشة الموضوعات، نقاط الضعف والقوة المرافقة لاستخدام مشاركين ذوي معرفة بلغتين في الاختبار لإثبات المعادلة بين نسخ الاختبار في لغات متعددة.

أما في الفصلين السادس والسابع الذي كتبته الدكتورة ليندا كوك. الدكتورة اليسيا لشميدت كاستكلار وكاترين براون (الفصل السابع فقط) من هيئة الاختبارات التربوية قدم وصفاً لمنهج مقارن إحصائية الاختبارات في لغات متعددة وعرضاً لموضوعات هامة قد تنتج عن ترجمة وتكييف تلك الاختبارات.

نأسف أن نقول إن اليسيا شميدت كاستكلار قد غادرت الحياة عام 2003. كانت محاضرة مدعوة في مؤتمر ITC في جورج تاون وكانت مساهمة مهمة في بحث منهجية الاختبارات، ومنها منهجية ترجمة وتكييف الاختبارات.

كان الهدف من الفصلين الثامن والرابع عشر تحويل التركيز عن تقديم الموضوعات الأولية والمنهجية إلى عالم التعقيدات في تطبيق ترجمة وتكييف الاختبارات.

إن تطبيق ترجمة الاختبارات ومنهجة التكييف تتضمن اختبارات معتمدة، اختبارات ذكاء، اختبارات معرفية، اختبارات تستعمل في إطار صناعي وتنظيمي، اختبارات قبول واختبارات شخصية. تصف الدكتورة سيندي فينزجرالد، مستشارة سابقة في مايكروسوفت وحالياً في كافيون، في الفصل الثامن عملية ترجمة وتكييف الاختبارات المعتمدة التي تستعملها مايكروسوفت.

إن استخدام النظم الموجودة على الإنترنت لتسهيل عمل المترجمين تبدو مثلاً جيداً في المهنة. في الفصل التاسع يصف الدكتور كارلوس مالدونادو (بوتنا/ وستشتر الشمالية، الولايات المتحدة) والبروفسور كيري كسينجر (جامعة سانت

توماس، الولايات المتحدة) مشكلات في الترجمة من الإسبانية إلى الإنجليزية وفي تكييف واحد من أكثر أدوات اختبار الذكاء شعبية في العالم: مقياس وفكسلر للذكاء للبالغين. في الفصل العاشر يقدم البروفسور نوربرت تانزر (الجامعة العالمية المتحدة/ الولايات المتحدة) وجامعة (غراز/ استراليا) مناقشة جديدة في التطور المتزامن لبعض الاختبارات النفسية عوضاً عن ترجمة وتكييف الاختبارات عبر اللغات والثقافات المختلفة.

في الفصل الحادي عشر يصف البرفسور فرتز دراسكو وتاهيرا بروست من جامعة إيلنوي/ الولايات المتحدة عملهما المهم في إحداث اختبارات متعادلة عبر مجموعات مختلفة اللغات والثقافات كانت تستعمل سابقاً في أطر صناعية وتنظيمية.

أما في الفصل الثاني عشر يصف الدكاترة ميشيل بيلر (من هيئة الاختبارات التربوية) نعومي غافني وبننا هناني (المعهد القومي للاختبارات والتقويم في إسرائيل) جهودهم الطموحة في تهيئة اختبارات قبول بست لغات تستعمل في إسرائيل. في الفصل الثالث عشر يقدم بيتر ميرندا (جامعة رود أيلند الولايات المتحدة) عدداً من ملاحظاته ونتائجه في حقل ترجمة وتكيف الاختبارات عبر ممارسته لذلك العمل. قام عدد من الباحثين بعمل أطول وأنجح في ذلك الحقل. أخيراً في الفصل الرابع عشر يزودنا البروفسور سبيلبيرغر من جامعة جنوب فلوريدا واثنان من زملائه مانولت موسكو وتوماس برنر من ذات الجامعة بمعلومات غنية عن الموضوعات والطرق المتعلقة بترجمة وتكييف اختبارات الشخصية.

بالنيابة عن نفسي وعن زملائي بالتحرير، بيتر ميريندا وتشارلز سبيلبيرغر نأمل أن تكون هذه المجموعة من الفصول الأربع عشرة تعزز مهمة الهيئة الدولية للاختبارات وذلك بتوفير إرشادات وتحفيز أبحاث على الموضوعات التي تزداد أهميتها في ترجمة وتكييف الاختبارات. عن التطور في هذا الحقل كان هائلاً منذ تساؤلاتي الأولى عام 1989 . اليوم تطورت الدراسات في هذا الحقل أكثر، وأصبحت

هناك خطوط عريضة موجودة وجرى تنظيم المنهجيات وامتدادها، وعدد متزايد من الأمثلة النموذجية للممارسين لاتباعها. في ذات الوقت هنالك كثير من الأبحاث التي يجب القيام بها ونأمل أن تحفز هذه المجموعة من الفصول العمل للتقدم في هذا المجال.

القسم الأول

تكييف الاختبارات التربوية والنفسية

عبر الثقافات: موضوعات نظرية ومنهجية

1

موضوعات وخطط، وإرشادات تقنية لتكييف الاختبارات للغات وثقافات متعددة

رونالد. ك هامبلتون

جامعة ماستشوست/ امهرست

يوجد اليوم عدد كبير من الإثباتات عن أن الحاجة إلى نسخ معدّة بلغات متعددة عن اختبارات الذكاء، الإنجازات، والشخصية. وتقارير مسح شاملة في ازدياد (مثال، اريسكان 2002، هامبلتون، 2002، هامبلتون دو يونغ، 2003، هاركنس، 1998). فعلى سبيل المثال، دعت الجمعية الدولية لتقييم الإنجازات التعليمية (IFA) الأبحاث التالية في الرياضيات والعلوم (TIMSS) العالمية في أكثر من 45 دولة، وكان من مهامه تهيئة اختبارات في الرياضيات والعلوم في أكثر من 30 لغة. هناك أمثلة بارزة لمشاريع جديدة في تكييف الاختبارات في الولايات المتحدة تتضمن خططاً لإعداد نسخ في اللغة الإسبانية لاختبار القبول الجامعي اختبار التقييم المدرسي، (SAT)، اختبار مجلس التعليم الأميركي "تطور التعليم العام (GED) واختيار الإدارة التعليمية في الولايات المتحدة "التقييم الوطني للتقدم التعليمي" وعدد كبير من اختبارات الإنجاز في إدارات التعليم الحكومية. وبالفعل من المتوقع القيام بعملية تكييف عدد أكثر من الاختبارات في المستقبل؛ لأن (أ) أصبح التبادل الدولي للاختبارات أكثر شيوعاً، (ب) تستعمل اختبارات أكثر للحصول على المصداقية الدولية، (ج) وازدياد الاهتمام بأبحاث عبر الثقافات.

بالرغم من أن أسباب تكييف الاختبارات من لغة وثقافة إلى أخرى واضحة فعلى سبيل المثال، تسهل دراسة مقارنة الانجازات المدرسية عبر مجموعات مختلفة الثقافة واللغة، توفير المال والوقت المتعلق بتهيئة اختبارات جديدة، والعدل في التقييم فإن الطرق والخطوط العريضة لإعداد تكييف الاختبارات ومعادلة النتائج ليست معروفة جيداً (هامبلتون، 1993، هوي وترياندس، 1985، فان دي فيفر وهاملتون، 1996). حتى إن بعض الباحثين في الدراسات عبر الثقافات علق أن نسبة كبيرة من الأبحاث في ذلك الحقل يحتوي كثيراً من الأخطاء إلى حد جعله غير صالح بسبب عملية التكييف السيئة.

إن القصد من هذا الفصل هو (أ) مراجعة عدد من المصادر عن الأخطاء وعدم الصدق المرافقة لتكييف الاختبارات واقتراح طرق لتقليل تلك الأخطاء (ب) وصف بعض الخطط لتكييف الاختبارات التي قامت بها الهيئة الدولية للاختبارات (ITC) بمساعدة سبع وكالات عالمية (هاملتون، 1994، فان دي فيفر وهاملتون، 1996).

قبل البدء يجب التمييز بين تكييف الاختبارات وترجمتها. يفضل استخدام «تكييف الاختبار» على المصطلح «ترجمة الاختبار» الذي هو أكثر شيوعاً واستعمالاً في هذا الفصل: لأن المصطلح الأول أوسع وأكثر انعكاساً على ما يجب القيام به في الواقع عند إعداد اختبار تم إعداده للاستخدام في لغة وثقافة واحدة للتطبيق في لغة وثقافة أخرى.

يتضمن تكييف الاختبار كل الأنشطة بدءاً من تقرير عما إذا كان باستطاعة الاختبار تقدير تركيبة الاختبار ذاتها في لغة وثقافة أخرى، اختبار المترجمين، تقرير التكييف المناسب الذي يجب القيام به لإعداد الاختبار للاستعمال في لغة ثانية، إلى تكيف الاختبار والتأكد من تطابقه مع الشكل المكيف، إن ترجمة الاختبار خطة واحدة من عملية تكييف الاختبار وحتى في تلك الحالة مصطلح التكييف مناسب أكثر من مصطلح الترجمة لوصف العملية الحقيقية التي تجري؛ ذلك لأن المترجمين يحاولون الحصول على مفاهيم، مفردات وتعابير متعادلة ثقافياً، نفسياً

ولغوياً للغة والثقافة الأخرى، بذلك تأخذ المهمة أبعاداً أكثر من ترجمة محتويات الاختبار حرفياً.

نستعمل المصطلح "اختبار" لغايتنا في إدراج كل النماذج والأدوات التربوية والنفسية، وحتى عمليات المسح والاستبيانات.

أسباب الأخطاء وعدم صدق تكييف الاختبار:

تؤمن الجمعية الأميركية للأبحاث التعليمية (AERA)، الجمعية النفسية الأميركية (APA)، والهيئة الوطنية للمقاييس في التعليم (NCME)، المعايير للاختبارات التربوية والنفسية (1985) تعليمات دقيقة للاختصاصيين في المقاييس التعليمية والنفسانيين الذين يختارون، يطورون، ويشرفون، ويستخدمون الاختبارات النفسية والتعليمية، هناك ثلاثة معايير في هذا الكتاب متعلقة بموضوع تكييف الاختبار.

المعايير 6.2 عندما يقوم مستخدم الاختبار بتغييرات أساسية في بنية الاختيار، طريقة الاستخدام، التعليمات، اللغة أو المحتوى، يجب عليه إعادة صدق استخدام الاختبار حسب حالات التغييرات أو عرض أسباب منطقية تدعم الادعاء أو مصداقية إضافة ليست ضرورة أو ممكنة.

المعايير 13.4 عندما يترجم اختبار من لغة/ لهجة إلى أخرى لاستخدامها لاختبار مجموعات ذات لغة واحدة يجب التثبت من مصداقيتها وجدارتها.

المعايير 13.6 إذا كان المقصود مقارنة نسختين لاختبارين في لغتين، يجب أن يدون دليل على مقارنة الاختبار.

توفر هذه المعايير خطوط عمل لاعتبار مصادر الأخطاء أو عدم الصدق الناتجة عن الجهود لتكييف الاختبار من لغة إلى أخرى ومن ثقافة إلى أخرى.

ولأغراضنا من الممكن تنظيم مصدر الأخطاء أو عدم صلاحيتها في ثلاث فئات عامة: (أ) اختلافات لغوية ثقافية، (ب) موضوعات تقنية، خطط، وطرق (ج) ترجمة

النتائج. إن الفشل في الاهتمام بمصادر الخطأ في كل من تلك الفئات يمكن أن ينتج عن عدم مساواة الاختبار الذي جرى تكييفه عند استخدامه في مجموعتين المختلفتين لغوياً وثقافياً. إن الاختبارات غير المتساوية، عندما يفترض أن تكون متساوية تؤدي إلى أخطاء في التفسير ونتائج مغلوطة عن المجموعات المشاركة.

مثال جيد عن خطأ في التفسير بسبب تكييف شيء لاختبار (هذا المثال قدمه ريتشارد وولف من كلية المعلمين في كولومبيا، وهو رائد في مهنته في حقل التقويم الدولي). وفي دراسة مقارنة دولية في القراءة (1990)، طلب من طلاب أميركيين دراسة أزواج من المفردات وتعريفهم كمتماثلين أو مختلفين في المعنى:

أدرجت الكلمتان "متشائم - دموي المزاج" ضمن مجموعة المفردات التي أحرز فيها الطلاب نقاطاً متوسطة (54٪ من الطلاب الأميركيين أعطوا الإجابة الصحيحة). كانت البلاد غير الناطقة بالإنجليزية الأولى في الأداء 98٪ من الطلاب أعطوا الإجابة الصحيحة. من خلال محاولة معرفة أسباب الاختلاف الكبير في الأداء اكتشف أن كلمة (دموي المزاج) ليس لها مرادف في ذلك البلد المرتفع الأداء ولذلك استعملت كلمة (متفائل)، جعل هذا التبديل السؤال أسهل وكان، من الممكن لنسبة كبيرة من الطلاب الأميركيين الإجابة عنه بالشكل الصحيح لو قدمت الكلمات المزدوجة بشكلها الجديد (متشائم - متفائل). إن الغرض من هذا المثال التركيز على الخطر من أخذ الاستنتاجات في الدراسات العالمية المقارنة للأداء دون دليل على عملية التكييف الناتجة في اختبارين متماثلين. قبل 1990 كان هناك كثير في المبادرات لدراسات عالمية تشمل أكثر من استخدام بعض المترجمين الجيدين، ويمكن مقارنة اختلاف ذلك مع تكييف الاختبارات المتطور الذي نشاهده اليوم في جمعية (TIMSS) ومنظمات التعاون الاقتصادي وبرامج التطوير للتقويم الدولي للطلاب (OECD)، (PISA انظر غريساي، 2003، هاملتون، 2002).

فيما يلي مناقشة عدة أغلاط شائعة وكيف يمكن معالجتها بشكل عملي.

الاختلافات اللغوية/ الثقافية التي تؤثر في النتائج:

إن تقويم وترجمة النتائج عبر الثقافات لا يجب أن ينظر إليها من الزاوية الضيقة لترجمة وتكييف الاختبارات (فان دي فيغر ولونغ، 1997، 2000). لكن يجب اعتبار هذه العملية ضمن كل مراحل عملية التقويم ومن ضمنها تساوي بنية الاختبار، إدارة الاختبار، بنية البنود المستعملة، وأثر السرعة على أداء الممتحن هذه العوامل الأربع سيجري مناقشتها لاحقاً، وستلقى اهتماماً أكثر في الفصول التالية:

البنية المتكافئة Construct Equivalence

يتضمن تكافؤ البنية كلاً من التكافؤ في المفهوم/ الوظيفة بالإضافة إلى التساوي في طريقة قياس البنية في عملية الاختبار في مجموعة مختلفة اللغة/ الثقافة (هاركنس، 1998). إذا فرضنا وجود البنية المتكافئة بين الثقافات المختلفة التي جرت دراستها فإن القيام بالمقارنة بين الدراسات عبر الدول. عبر الثقافات وعبر اللغات أساسي. إن استخدام اختبار غير متكافئ البنية هو أكثر الأخطاء أهمية في البحث عبر اللغات المختلفة.

على سبيل المثال: مقارنة أداء دولتين في الرياضيات: إذا كان اختبار المحتوى يعكس الاهتمام الأكبر للرياضيات في المناهج الدراسية في دولة وليس بذات الأهمية في الدولة الأخرى. مثال آخر: يمكن أن يكون في بنية (نوعية الحياة) يمكن أن يتضمن مفهوم الحياة كثيراً من الأمور المادية كالسيارات، المنازل، التلفزيونات، بينما لا يتضمن مفهوم الحياة في الدولة الأخرى أكثر من الطعام للبقاء وطبيب قريب من المنزل. إن مقارنة النتائج في اختبار الدولة ذات نوعية حياة جيدة وتم تكييفه للاستخدام في الأخرى له قيمة ضعيفة.

فإذا قررنا عما إذا وجود البنية المتكافئة بين ثقافتين يتضمن استراتيجية عقلانية يجب أن يبدأ الباحث باستخدام بديهته للإجابة عن أسئلة عديدة، على سبيل المثال هل من المعقول مقارنة تلك الثقافتين حسب تلك البنية؟

هل تلك البنية الذي تم دراستها لها معنى مواز في كل الثقافات التي يجري مقارنتها؟ هل تلك البنية فعالة في تك الدراسات.

لكي نستطيع الإجابة بنعم عن تلك الأسئلة وضمان تكافؤ المفاهيم/ الوظيفة وتكافؤ فعالية تلك البنية يجب اتخاذ عدة طرق. من الممكن القيام بهذا عن طريق مقابلة وملاحظة الأشخاص في الثقافات المعنية، إجراء الأبحاث عن تلك الثقافات وطرح أسئلة على آخرين يعرفون تلك الثقافات. إن هذه الطرق موضوعية ولذلك فإن استخدام مصادر أدلة مختلفة مستحسن جداً. فان دي فيفر وبورتينغا (الفصل الرابع) لديهم الكثير ليقولوه في ذلك الفصل عن الحكم على البنية المتكافئة.

إدارة الاختبار:

تهدد صعوبات التفاهم بين الذين يجرون الاختبار وبين الذين يديرون الاختبار صدق نتائج الاختبار بشكل كبير. من الممكن أن تكون تعليمات الاختبار غير واضحة بسبب صعوبات في الترجمة. إحدى الطرق للحيلولة دون ذلك، وهي ممكنة، هو التأكد من كون التعليمات في الاختبار نفسه واضحة ومفهومة بذاتها وبأقل اعتماد على الاتصال اللفظي (فان دي فيفر وبورتينغا، 1991).

ومن المتوقع وجود بعض الصعوبات الأخرى في تعليمات تقدير مقياس الدرجات المستخدم في "قياس الموقف" أيضاً؛ لأن تلك الاختبارات ليس لها وجود في الدول الأخرى (انظر هاركنس، 1998).

إن اختيار الإداريين المناسبين للاختبار من الممكن أن يكون مفيداً أيضاً. فيجب أن تتوفر بينهم الشروط التالية (أ) أن يجري اختبارهم في موطن المجموعة التي تجري الاختبار (ب) أن يكونوا مطلعين على الثقافة، واللغة، واللهجة (ج) لديهم مهارات كافية في إدارة الاختبارات (د) أن يدركوا أهمية تطبيق الإجراءات المتبعة أثناء الاختبار.

بالإضافة إلى ذلك فإن التناسق في إدارة الاختبارات لمجموعات مختلفة يمكن أن يكون أفضل إذا توفر التدريب الأساسي لكافة الأشخاص الذي يديرون الاختبار.

يجب أن يخطط لحصص تدريبات كجزء من عملية تطوير الاختبار والتأكيد على وضوح وعدم غموض التواصل بين الإداريين والممتحنين، أهمية اتباع إرشادات، ضبط الوقت المحدد للاختبار، وتأثير مقدمي الاختبار على جدارة وصدق الاختبار وغير ذلك.

شكل ومحتوى الاختبار Test Format

إن التفاوت المألوف في بنود البنية يشكل مصدراً آخر لعدم مصداقية نتائج الاختبار في الدراسة عبر الثقافات. في الولايات المتحدة استخدمت بنود استجابة متعددة مثل بنود ذات عدة إجابات لاختيار أحدها بشكل كبير في الاختبار (مع أن هذا يجري تغييره في السنوات العشر الأخيرة واليوم نرى استعمال أسئلة تقييم الأداء أكثر). في دراسة عبر الثقافات لا نستطيع الجزم أن كل المتقدمين للاختبار مطلعين على طريقة تلك الأسئلة مثل الطلاب الاميركيين. إن الدول التي تتبع النظام البريطاني في التعليم (تاريخياً على الأقل) تؤكد أكثر على كتابة المقالات وأسئلة ذات إجابة قصيرة بالمقارنة مع الأسئلة ذات الإجابات المتعددة. وبذلك يكون الطلاب من تلك البلاد في وضع خاسر مقارنة مع نظرائهم الأميركيين. عندما يكون التأكيد في بنية الاختبارات على استجابات مثل كتابة المقالات كطريق أساسية للتقويم يكون الممتحنون ذوو المعرفة بالأسئلة ذات الإجابات المتعددة في وضع صعب. في بعض الأحيان يكون التوازن في بنية الاختبار الأفضل لتحقيق العدالة والتقليل من عدم مصداقية عملية التقويم. وقد اعتمدت هذه الطريقة في الدراسة الدولية الحديثة للأداء (TIMSS & OECD / PISA).

هناك حل آخر للتغلب على أثر احتمال التحيز المرافق لبنية أحد البنود بشكل خاص هي في أن يتضمن الاختبار الذي يقوم بتقييم المجموعات البنود المألوفة لكل تلك المجموعات. عندما يكون من المؤكد أن الطلاب ليسوا في وضع صعب وأن كافة المعطيات جرى قياسها من المفضل استخدام الأسئلة ذات الإجابات المتعددة أو مقياس درجات بسيط.

إن الميزة الكبرى لاستخدام الأسئلة ذات الإجابات المتعددة أو مقياس درجات بسيط هو أن تقديرها أكثر موضوعية وبذلك يمكن تجنب تعقيدات المقاييس التي تصاحب الإجابات ذات النهاية المفتوحة، هذا يرتبط بشكل خاص في الدراسات عبر الثقافات، حيث من الممكن أن يكون ترجمة قواعد المقاييس أكثر صعوبة من الاختبار نفسه. بالإضافة إلى ذلك فإن تعليمات كثيرة واضحة تتضمن أمثلة وتمارين تساعد في الإقلال من تفاوت الاعتياد (فان دي فيفر وبورتينغا، 1992). في ذات الوقت فإن استخدام بنية بند واحد للاختبار قد يكون من الخطورة أن يضيق بنية الاختبار المهمة إلى تلك الأجزاء التي يمكن قياسها حسب بنية البند الأوحد، أيضاً قد يحرف نتائج الدراسات المقارنة عبر الحدود القومية.

السرعة:

جرى غالباً الافتراض أن الممتحنين يعملون بشكل سريع في الاختبارات السريعة (فان دي فيفر وبورتينغا 1991) ولكن معرفة العمل السريع هي مهارة في عملية أخذ الاختبار التي من الممكن أن لا تكون معروفة أو مفهومة من قبل الممتحنين في ثقافات مختلفة في دراسة تُقارن بين طلاب هولنديين وطلاب وعرقيات مختلفة في هولندا، وجد فان ليست وبريخرودت (1995) أن عامل السرعة ضاعف الانحياز في الدرجات.

لأنه ليس، لدى كل الثقافات الخبرات في الاختبارات السريعة وكان الممتحنون الفاقدون لتلك الخبرة في وضع حرج جداً. هنالك كثير من الدراسات المختلفة التي تسلط الضوء على بنود وتميز في الاختبار بسبب دور السرعة في الاختبار (انظر دراسات حول التميز العرقي في اختبارات تقويم الطلاب في الولايات المتحدة). على سبيل المثال في (SATs) بعض البنود الأخيرة في الاختبار عادة تظهر تحيزاً أكثر من تلك التي توجد في بداية الاختبار. يكون التمييز هذا ضد القراء الضعفاء وهذا غالباً ما يكون بسبب دور السرعة في أداء الاختبار. إن الحل الأمثل هو التقليل من

السرعة كعامل في اختبار أداء المعرفة إلا إذا كانت جزءاً من البنود التي يجري قياسها. إن النقطة الأخيرة مهمة جداً لأنه في بعض الأحيان تكون السرعة في الأداء جزءاً متكاملاً من قياس بنية الاختبار كما في حالة القدرة على حل مسائل في التحليل الاستنتاجي. عندئذ تكون السرعة قسماً مهماً من الاختبار وعلى الممتحنين فهم ضرورة العمل السريع.

موضوعات تقنية، خطط، وطرق:

هنالك خمسة عوامل تقنية تؤثر في صدق الاختبارات المكيفة للاستخدام في لغات وثقافات أخرى. الاختبار نفسه، اختيار وتدريب المترجمين، عملية الترجمة، خطط عقلانية لتكييف الاختبار وخطط لجمع المعطيات وتحليلها لتثبيت التكافؤ. سيجري بحث كل من هذه العوامل بشكل مختصر. وتظهر مناقشة تلك العوامل بشكل مفصل في الفصول اللاحقة.

الاختبار:

إذا كان الباحث يعرف أنه سيستخدم الاختبار للغة أو ثقافة مختلفة، المفيد أن يضع ذلك في الحسبان في بداية عملية تطوير الاختبار. وإذا أخفق في ذلك فسينتج عن ذلك صعوبات في عملية التكييف التي تؤدي بدورها إلى تخفيض صدق الاختبار المكيف (هامبلتون وباتسولا، 1999).

إن اختيار شكل الاختبار من مواد محفزة له، المفردات، تركيب الجمل ونواح أخرى يمكن أن تشكل صعوبة في الترجمة الجيدة التي يجب أن تؤخذ جميعها بالحسبان عند إعداد مواصفات الاختبار.

يمكن لذلك العمل الوقائي التقليل من المشكلات اللاحقة. على سبيل المثال أسئلة عن النقود يمكن حذفها لأن العملات مختلفة في العالم ومن الممكن صعوبة إيجاد تكافؤ في ترجمتها لوضعها في الاختبار. كذلك نص القراءة عن موضوعات خاصة بإحدى الثقافات مثل "الهوكي" تبدو غير مألوفة في عدة ثقافات أخرى

ويمكن رفضها والاستعانة بمقاطع عن المشي في الحديقة أو نشاطات أخرى يمكن أن يكون لها معنى في لغات وثقافات مجموعات أخرى. وتنشأ صعوبة أخرى في تكييف النصوص من الإنجليزية إلى لغات أخرى وهي وجود "المبني للمجهول" في النص لأن هذا الزمن في القواعد موجود في اللغة الإنجليزية ولكنه غير موجود في لغات أخرى (الإسبانية على سبيل المثال).

أما في معيار الشخصية، فيجب أن يؤخذ الحذر في اختيار المواقف، المفردات، والتعابير التي يمكن تكييفها بسهولة عبر الثقافات/ اللغات المختلفة للمجموعات. على سبيل المثال: يمكن أن يكون بعض أنواع السلوك عادياً في العالم الغربي ولكن له معنى آخر أو ليس له أي معنى في ثقافات أخرى. عبارة "أحب أن أقوم بالمحادثة في حفلة" ليس لها معنى في ثقافة لا يكون فيها حفلات أو حيث لا تذهب النساء إلى الحفلات أو حيث المبادرة إلى الحديث يمكن أن يكون تصرفاً غير مقبول. هذا واحد فقط من الأمثلة التي من الممكن مواجهتها.

اختيار وتدريب المترجمين:

إن أهمية الحصول على خدمات مترجمين مؤهلين واضحة. حاول الباحثون متابعة عملية الترجمة لمترجم واحد تم اختياره لأنه كان من الممكن الوصول إليه/ إليها لكونه صديقاً، زوجة، أو شخصاً يمكن استخدامه بمبلغ بسيط إلى ما هنالك. إن عمل الترجمة الكفء لا يمكن أن يعتبر أمراً مفروغاً منه، كذلك فإن استخدام مترجم واحد مؤهل أو غير مؤهل لا يسمح بالحصول على تفاعل ذي قيمة بين المترجمين المختلفين لإيجاد الحلول لنقاط عديدة تنشأ عند القيام بعملية تكييف الاختبار. قد يكون لأحد المترجمين على سبيل المثال وجهة نظر في استخدام مفردات أو تعابير مفضلة لديه قد لا تكون مناسبة لتحقيق تكييف جيد للاختبار. من الممكن أن يكون استخدام مترجمين عدة حماية ضد أخطار استخدام المترجم الوحيد مع تفصيلاته وخصوصيته اللغوية.

في ذات الوقت، يجب أن يكون المترجمون أكثر من أشخاص مؤهلين ومتآلفين مع اللغات المستخدمة في الترجمة، يجب أن يعرفوا الثقافات جيداً وبشكل خاص الثقافة التي يترجم إليها (الثقافة المرتبطة باللغة التي يجري ترجمتها). إن هذه المعرفة أساسية في فعالية التكييف. كذلك من المفضل جداً معرفة مواد الموضوعات في تكييف اختبارات الأداء. إن دقة وفروق المعنى في موضوع ستخفى عن مترجم ليس له معرفة في ذلك الموضوع وغالباً يعود المترجمون الذين يجهلون المعرفة التقنية إلى الترجمة الحرفية التي تؤدي بدورها إلى إحداث صعوبات للطلاب الذين يجرون الامتحان في اللغة الثانية وتهدد صدق الاختبار. مثلاً إن جملة "Je ne suis pas une valise" في الفرنسية لها ترجمة حرفية سهلة في الإنجليزية، (أنا لست حقيبة) ولكن المعنى الحقيقي لتلك الجملة في اللغة الفرنسية هو "لست غبياً إلى هذا الحد" إن الترجمة الحرفية من الفرنسية إلى الإنجليزية قد شوه المعنى بالكامل.

أخيراً إن المترجمين سوف يستفيدون من بعض التدريب في تكوين الاختبار. مثلاً يجب أن يعرف المترجمون عندما يقومون بتكييف اختبارات الأداء والأهلية عدم إحداث قوافي لغوية تقود الممتحنين إلى الإجابات الصحيحة وترجمة البنود المتشابهة في الأسئلة ذات الإجابات المتعددة بشكل يجعلها ذات معنى واحد. إن المترجم الذي ليست لديه أي معرفة في مبادئ الاختبار وبناء المقاييس يمكن بسهولة أن يجعل الاختبار أقل أو أكثر صعوبة بدون قصد وذلك بدوره يؤدي إلى عدم صدق الاختبار في المجموعة المستهدفة.

عملية الترجمة:

قد تهدد اللهجات في لغة ما صدق تكييف الاختبارات، أي لهجة هي الأهم أو هي الهدف المستخدم في التكييف الذي يمكن تطبيقه داخل اللغة الواحدة؟ يجب البت في هذه المسألة قبل البدء في تكييف الاختبار ويجب استخدامها ضمن المواد

لتدريب المترجمين. إن إحصاء تكرار الكلمات قد يكون قيماً في الحصول على ترجمة اختبار صالح. على وجه العموم من الأفضل ترجمة الكلمات وتعابير مكونة من عدة كلمات بذات التواتر في اللغتين وذلك لمحاولة السيطرة على الصعوبات عبر اللغات. إن المشكلة هي أن لوائح تواتر الكلمات والتعابير ليست متوفرة دائماً وهذا سبب آخر لتفضيل المترجمين الذين لهم اطلاعهم الكامل على كلتا الثقافتين المصدرة والمستهدفة وليس معرفة اللغتين فقط.

تستعمل اللامركزية في بعض الأحيان في تكييف الاختبارات. من الممكن أن لا يكون لبعض الكلمات أو التعابير مرادف في اللغة المستهدفة. حتى إنه من الممكن أن لا توجد تلك الكلمات أو التعابير في تلك اللغة. إن عملية اللامركزية تتضمن مراجعة اللغة الأصلية المترجم منها الاختبار وبذلك يتم استخدام أساس لغوي مترادف في لغة لنسختين المصدر والمستهدفة. إن اللامركزية ممكنة عندما يكون الاختبار الأصلي في مرحلة التحضير في ذات الوقت الذي يتم فيه إنجاز نسخة اللغة المستهدفة. وهذا يكون موجوداً عند إعداد اختبارات التقويم العالمية. وبعض الاختبارات المعتمدة (الاختبارات المعدة من مايكروسوفت) التي أعدت للاستخدام في العالم.

خطط الحكم النقدي لتكييف الاختبارات:

إن الخطتين المفضّلتين في الترجمة هما الترجمة المبكرة والترجمة الراجعة، إن خطة الترجمة المبكرة هي أن مترجماً واحداً أو من الأفضل عدة مترجمين يقومون بتكييف الاختبار من لغة المصدر إلى اللغة المستهدفة. عندئذ يجري الحكم على تعادل النسختين المترجمتين من الاختبار من قبل مجموعة ثانية من المترجمين. يمكن إجراء مراجعة على نسخة الاختبار في اللغة المستهدفة لتصحيح بعض الأخطاء التي وجدها الفريق الثاني من المترجمين. في بعض الأحيان وكخطوة أخيرة يقوم شخص ثالث ليس بالضرورة أن يكون مترجماً بتحرير الاختبار بجعل اللغة أكثر سلاسة لأنه

في بعض الأحيان يحصل بعض التفكك في اللغة أثناء الترجمة، التي يقوم بها عدة مترجمين أو مجموعات للنسخة الواحدة.

إن الميزة الأساسية لخطة الترجمة المبكرة هو أن الحكم يصدر مباشرة على النسخة الأصلية من الاختبار والنسخة المترجمة. إن صدق الحكم على تكافؤ النسختين يُعزز بوجود مجموعة صغيرة من الممتحنين ليزودوا المترجمين بملاحظاتهم عن الاختبار والإرشادات، المحتوى أو الشكل العام من الممكن القيام بذلك في دراسات تدعى "فكر بصوت عال".

أما نقطة الضعف الأساسية في خطة الترجمة المبكرة فهي مرتبطة مع المستوى العالي من الاستنتاجات التي يقوم بها المترجمون عن التكافؤ بين نسختي الاختبار. هناك نقاط ضعف أخرى مثل (أ) قد يكون لدى المترجم مهارة في إحدى اللغات أكثر من الأخرى. (ب) أن الحكم على تكافؤ الاختبار يقوم به أشخاص ثنائيو اللغة وبهذا يمكن أن تكون نظرتهم التخمينية متركزة على معرفتهم لكلتا اللغتين، (ج) أن المترجمين قد يكونون متعلمين أكثر من الطلاب ذوي اللغة الواحدة الذين يتقدمون للاختبار وبذلك يخفق المترجمون في إدراك بعض الصعوبات التي تواجه الممتحنين، (د) أن الأشخاص الذين يطورون الاختبار ليسوا في موقع يستطيعون به أن يصدروا أحكاماً عليه بأنفسهم.

إن خطة الترجمة الراجعة هي المعروفة والأكثر شيوعاً في حفظ الحكم النقدي للاختبارات. في نسختها الأكثر شيوعاً، يقوم واحد أو أكثر من المترجمين بتكييف اختبار من اللغة الأصلية إلى اللغة المستهدفة، ثم يقوم مترجمون مختلفون بترجمة الاختبار إلى اللغة الأصلية. ويجري مقارنة النسختين، الأصلية والمعادة الترجمة ويجري تقويم التكافؤ بينهما. إذا كانت النسختان متشابهتين يجري الموافقة على التكافؤ بينهما، إن خطة الترجمة الراجعة يمكن استخدامها لاختبار عام لكل من نوعية الترجمة ولكشف بعض المشكلات التي ترافق عملية ترجمة أو تكييف غير

جيدة، يفضل الباحثون تلك الطريقة بشكل خاص لأنها تزودهم بفرصة للحكم على النسختين الأصلية والمترجمة للاختبار وبذلك يستطيعون تكوين رأيهم الشخصي عن عملية التكييف. وهذا ليس ممكناً في الترجمة المبكرة إلا إذا كانت لهم المهارة في اللغتين.

بالرغم من أن الترجمة الراجعة لها فضائلها وتستطيع تعيين المشكلات في عملية التكييف، لكنها نادراً ما تستطيع توفير الدليل الكافي لدعم صدق استخدام الاختبار المكيف. إن الدليل على أن تكافؤ الاختبار الذي توفره خطة الترجمة الراجعة هو واحد فقط من عدة أنواع من الأدلة التي يجري تصنيفها في دراسة تكييف الاختبار. أحد مواطن الضعف هو أن المقارنة بين نسخ اختبار في لغتين أو أكثر تجري في اللغة الأصلية فقط. من الممكن أن يكون تكييف الاختبار ليس جيداً بالرغم من أن دليل مقارنة الاختبار الأصلي واختبار الترجمة الراجعة يدل على غير ذلك. قد يحدث ذلك إذا استخدم المترجمون جميعاً قواعد واحدة للتأكد من أن الاختبار المترجم متشابه مع الاختبار الأصلي. هنالك نقطة ضعف أخرى هي أن كون التكييف ضعيف يعود إلى احتفاظه بأوجه غير مناسبة من اختبار اللغة الأصلي مثل بنية القواعد الواحدة والتهجئة. قد تسهل تلك الأخطاء عملية الترجمة الراجعة ولكن تلك الخطة قد تخفي نقاط ضعف مهمة في نسخة اختبار اللغة المستهدفة على سبيل المثال من الممكن الاحتفاظ بكلمة "Icehocky" عند ترجمة اختبار إلى الإسبانية وعندئذ يكون من السهل القيام بالترجمة الراجعة.

لسوء الحظ يمكن أن تكون تلك الرياضة لا معنى لها لكثير من الأشخاص الذين يتكلمون الإسبانية فقط وبذلك من الممكن حصول تدني صدق نسخة الاختبار باللغة الإسبانية.

أخيراً بالإضافة إلى ما تقدم، فإن خطط حكم نقدي أخرى لها بعض العوائق لأن بعض النماذج من المجموعة التي يجرى عليها الاختبار لا تقوم بالاختبار حقيقة

في ظروف اختبارية صحيحة (أو أي ظروف أخرى). هنالك دليل متوافر يوحي أن الذين يقومون بمراجعة الاختبار غير قادرين على تعيين الأخطاء في بنود الاختبار ولذلك يجرى اختبار ميداني بشكل دوري لبنود الاختبار قبل استخدامها. يجب أن تاختبار الاختبارات المكييفة ميدانياً أيضاً لكشف المشكلات التي لم يكتشفها المترجمون حتى عند استخدام مترجمين جيدين وخطط الترجمة المثلى معاً (انظر هاميلتون وبانسولا، 1999).

مخططات جمع المعطيات وتحليلها لإقامة تكافؤ البنود والاختبار:

هنالك ثلاث خطط شائعة الاستعمال في جمع المعطيات لتقويم التكافؤ في بنية الاختبار وبنوده في لغات مختلفة، فيما يلي تقويم تلك الخطط:

1- إجراء الطلاب "الذين يجرون الاختبار" الاختبار في اللغتين اللغة الأصلية واللغة المستهدفة. إن ميزة تلك الخطة أنه من الممكن ضبط الاختلاف في ميزات الطلاب يمكن جمع بنود وإحصائيات مختلفة عن (الاختبار من إداريين كل نسخة اختبار ومقارنتها لتقرير التكافؤ. على كل تعتمد هذه الخطة على الافتراض أن الطلاب ثنائيي اللغة لديهم مهارة متساوية في كلتا اللغتين. هذا لا يحدث غالباً في مجموعة كبيرة من الطلاب (تشيكو، 1987، روسانسكي، 1979) وبذلك يجب ملاحظة ذلك الافتراض حسب الإمكان.

لجعل خطة جمع المعطيات ثنائية اللغة صالحة من الأفضل تطبيقها مع خطة جمع معطيات ثانية وبذلك يمكن تقصي صدق متقارب للنتائج.

إن المشكلة الثانية في خطة جمع المعطيات هي أن نتائج الاحصائيات التي تم الحصول عليها من جمع المعطيات لا يمكن تعميمها ضمن مجموعة الطلاب ذوي اللغة الواحدة؛ لأن مجموعة ثنائيي اللغة (بشكل عام) مختلفين في عدة طرق عن نظرائهم (الطلاب ذي اللغة الواحدة) (هامبلتون، 1993).

في إحدى الدراسات التي أجراها هلن، دراسكو وكوموكار (1982) في "الدليل الوصفي الوظيفي" اكتشف هؤلاء الباحثون أن 4٪ من البنود في مقياس المواقف تم

تصنيفها على أنها ترجمة بشكل سيئ في أحد النماذج للطلاب ثنائيي اللغة. وتم تصنيف 30٪ من البنود على أنها ترجمة سيئة عندما استخدمها طلاب ذوو لغة واحدة (لغة المصدر واللغة المستهدفة).

هنالك خطة مختلفة عن الخطة الثنائية اللغة، التي لها ذات الحدود ولكنها أسهل في التطبيق، تتضمن طلاباً ثنائيي اللغة تم اختيارهم عشوائياً لأخذ أحد الاختبارات في تلك الحالة تظهر فاعلية تكافؤ خطة اختبار المجموعة العشوائي.

2- أخذ طلاب أحاديو اللغة الاختبارين، الاختبار الأصلي والاختبار ذا الترجمة الراجعة.

تتضمن هذه الخطة الإشراف على إجراء اختبار لغة المصدر والاختبار ذي الترجمة الراجعة على الطلاب أحاديي اللغة (لغة المصدر). يتم التعرف على تكافؤ البنود بمقارنة أداء المشاركين في كل من الاختبارين في كل بند. يمكن استخدام عملية تحليل العوامل على المعطيات المجموعة من كل اختبار ومقارنة بناء تلك العوامل. إن ميزة تلك الخطة هي أنه باستخدام نموذج واحد من المشاركين لا يكون هناك أي خلط في النتائج بسبب صفات الطلاب (هامبلتون وبولورك، 1991).

هنالك عيبان رئيسان يضعفان فائدة استخدام خطة جمع المعطيات:

أولاً : لا تجمع معطيات تجريبية من اختبار اللغة المستهدفة بمعنى أنه لا يستخدم طلاب أحاديو اللغة في اللغة المستهدفة مع أن الهدف من البحث هو تطبيق النتائج على نسخة الاختبار باللغة المستهدفة وعلى طلاب اللغة المستهدفة أحاديي اللغة.

ثانياً: لا تكون النتائج التي حصل عليها مستقلة لأنه لا يمكن استبعاد نتائج التعليم من الإشراف على الاختبار الأول في لغة المصدر الأصلية ولا تأثير التعليم على أداء الطلاب في اختبار الترجمة الراجعة. يستطيع التوازن تخفيض أهمية تأثير التطبيق ولكنها تصعب التحليل.

3- يأخذ طلاب لغة المصدر أحاديو اللغة اختباراً في لغة المصدر، ويأخذ طلاب اللغة المستهدفة أحاديو اللغة الاختبار في اللغة المستهدفة.

إن خطة جمع معطيات مناسبة تكون بأن يأخذ طلاب أحاديو اللغة اختبار لغة المصدر وأن يأخذ نموذج ثان من الطلاب الأحاديي اللغة اختيار اللغة المستهدفة. وعادة لا يمكن الاحتفاظ بافتراض مساواة توزع الإمكانيات بين المجموعتين. لحسن الحظ لا توجد ضرورة لهذا الافتراض إذا جرى القيام بالتحليلات حسب نظرية الإجابة عن كل بند (IRT) هيكلية الاختبار (إليس، 1989، 1991، اليس وكيميل، 1992، هامبلتون، سواميناثن وروجرزر 1991، فان دي فيفر وليونغ، 1997، 2000) أو إذا جرى القيام بالتحليلات حسب دراسات التكافؤ باستخدام ربط الإجراءات (هولاند ووينر، 1993). إن ميزة تلك الخطة هي أن نماذج المصدر والمجموعة المستهدفة قد جرى استخدامها في التحليل، وبذلك تكون نتائج تكافؤ الاختبار في اللغتين عامة في تلك المجموعات.

إن أحد أهم الاستقصاءات لترشيح إنجاز تكافؤ البنود هي دراسات البنود الموجه (هامبلتون وغيره، 1991، سيرغي وآلالوف، 2003). إن مقارنة إحصائيات البنود في اختبارات اللغتين (أو أكثر) كانت هي المسيطرة على أي فروقات الاستطاعة في المجموعتين (انظر هامبلتون وكانجي، 1995) إن البنود التي تحتوي الفروقات قد عُرفت ودرست بدقة لتحديد التعليل الممكن لتلك الاختلافات (انظر أركيكان، 2002).

إحدى تلك التعليلات هو التكييف السيئ. لسوء الحظ فإن هذه الدراسات غير قادرة على فصل الفروقات الثقافية عن مشكلات التكييف ولكن في أكثر الأحيان يكشف بشكل عام عن مشكلات محتملة في الاختبار المكيف. إن تحليل البنود الموجه تنتج عن نظريات اختبارات قديمة وجديدة ويمكن استخدامها في كل من استجابة المعطيات الثنائية ومعطيات استجابة ملحقة (سيرغي وآلالوف، 2003).

عوامل تؤثر في شرح النتائج:

في دراسات عبر الثقافات على مقياس عالمي، فإن الهدف من الاختبار هو أن يؤمن الأساس لإجراء المقارنات بين مجموعات مختلفة الثقافات واللغات لكي نستطيع فهم المفارقات والتشابه الموجودة (هاميلتون 1990، 2002). في بعض الأحيان يكون الاهتمام في المتغيرات المتشابهة وفي أحيان أخرى يكون التركيز على تقييم متغيرات الشخصية أو على معلومات عامة (نوعية الحياة، الصحة).

تأمل تلك الدراسات أن تستخدم تلك النتائج في البحث عن طرق لمقارنة المجموعات وفهم المفارقات بينها. لا يجب أن تستعمل الدراسات عبر الثقافات لدعم المناقشات عن التفوق الفريد لقومية وكأن دراسة الفروقات الدولية تعادل سباق جياد فيها الرابحون والخاسرون (وستبيري، 1992). في أحسن الحالات توفر تلك الدراسات لمحة عن المفارقات الموجودة، وتؤمن فقط أساساً محدداً لتفسير النتائج في هذا المفهوم لكسب فهم أكثر عند تفسير النقاط، ويجب أن نأخذ بعين الاعتبار عوامل أخرى خارجية ليس لها علاقة بالاختبار أو تقدير المعايير خاصة بجنسية محددة ، المناهج الدراسية، المستويات والسياسة التعليمية، الغنى، مستوى الحياة، القيم الثقافية إلى ما هنالك. من الممكن أن يكون كل ذلك عوامل أساسية في تفسير درجات صحيح عبر الثقافات/ اللغات ومجموعات دولية. من الطبيعي أن تناقش عينات من العوامل التي يجب التفكير بها عند تفسير نتائج الاختبار لمجموعات عبر اللغات والثقافة لاحقاً.

التشابه في المناهج الدراسية:

في نطاق وجود اختلافات في المناهج الدراسية، فإن مقارنة الأداء بين ثقافات مختلفة ستكون غامضة إذا لم تؤخذ تلك اختلافات بعين الاعتبار، لاحظ وستبيري (1992) أن نتائج "الدراسة العالمية الثانية للرياضيات" (SIMS) تشير إلى أن أداء الطلاب الأميركيين كان ضعيفاً في كل المراحل في كل مادة في الرياضيات التي

جرى تغطيتها في الاختبار. عند مقارنة أداء الطلاب اليابانيين والأمريكيين لوحظت فروق كبيرة في المناهج الدراسية في البلدين. على كل حال في نطاق المنهاج الدراسي المتشابه، لاحظ وسبيري أنه ليس هناك أي اختلاف في أداء الطلاب في البلدين.

إن أهمية تحليل الاختلافات في المناهج الدراسية واضح في دراسات المقارنة الدولية للأداء، ولذلك وبالرغم من كل المعارضات (بسبب الجهد والتكلفة) صنفت معطيات استبيان مكثفة مع معطيات الاختبار في كل دولة مشاركة.

دوافع الطلاب:

تساءل وينر عما إذا كان يمكن فصل الخبرة الظاهرة المقاسة في الاختبار عن الدوافع. لاحظ أن كل الطلاب الذين تم اختيارهم (بشكل عشوائي) للمشاركة في اختبار "دراسة التقويم الدولي للتقدم التعليمي" في كل دولة قد شعروا بالفخر لأنه قد تم اختيارهم لتمثيل مدارسهم ودولتهم، وبذلك أُلقيت عليهم مسؤولية الأداء الأفضل. وفي الجانب الآخر كانت المشاركة في الدراسة المقارنة العالمية للطلاب في دولة أخرى عبارة عن نشاط آخر ليس بتلك الأهمية لأن درجاتهم لم تكن متيسرة. كان الاختبار لهؤلاء الطلاب "رهاناً ضعيفاً".

إن تفسير اختلافات الأداء بين دول ذات طلاب ذوي دوافع ودول ليس لطلابها أي دوافع دون أي اعتبار لمتغيرات الدوافع في أداء الاختبار سينتج عنها إساءة تفسير خطيرة للنتائج.

العوامل السياسية الاجتماعية:

إن معنى وتفسير الدرجات تختلف حتى وإن كانت الدرجات متشابهة، فكر في إجراء مقارنة درجات الاختبار بين طلاب من دول متطورة ودول نامية أو مجتمع صناعي ومجتمع ريفي. في ذلك المحيط، لا يكون أداء الطلاب ذا علاقة باستطاعتهم على الإطلاق، بل قد يكون الأداء انعكاساً لعدم القدرة للحصول على مصادر كافية أو النوعيات المختلفة للخدمات التعليمية المتاحة.

إن النقطة الأساسية هي أنه للحصول على تفسير ذي دلالة للنتائج يجب حسبان الحقائق الاجتماعية، السياسة والاقتصادية المختلفة التي تواجه الأمم كما يجب حساب الفرص التعليمية المتاحة في ظل تلك الحقائق (اولدا، 1981). لذلك من المهم أن يكون المسؤولون عن تطوير الاختبارات وصناعة سياستها مطلعين على تلك الموضوعات الثقافية بحد ذاتها والتي يمكن أن تؤثر في أداء الاختبار.

إرشادات عملية لتكييف الاختبارات:

من المؤكد أن الكتابات التقنية لتوجيه عملية التكييف غير كاملة (من وجهة النظر التقديرية) ومتفرقة في كثير من المطبوعات العالمية، التقارير، والكتب. لم يكن هناك أي مصدر كامل يستطيع الممارسون الرجوع إليه لبعض النصائح ولم تتكون أي مجموعة من الإرشادات لتكييف الاختبار (هامبلتون، 1994، فان دي فيفر وهامبلتون، 1996) ولم تكن طرق القياسات المعقدة (نماذج بنود الاستجابة ونماذج توازن التركيبات) التي تساعد في إنشاء عملية معادلة الدرجات التي تم الحصول عليها في الاختبارات المكيفة لاستخدامها في اللغات والثقافات المختلفة معروفة للباحثين الذين يقومون بتكييف الاختبارات حتى وقت قريب (هولن، 1987). لكن كما هو واضح في فصول هذا الكتاب (انظر هامبلتون ودي جونغ، 29003) فإن الوضع قد تحسن جوهرياً منذ أوائل التسعينيات. في الواقع كان الهدفان من مؤتمر ITC الذي عقد في جورج تاون في الولايات المتحدة عام 1999 أولاً: جمع باحثين من العالم لتبادل المعرفة والخبرة في تكييف الاختبار. وثانياً: تقديم النسخة النهائية للدليل الموجز لتكييف الاختبارات الذي أشرفت عليه الهيئة العالمية للاختبارات (ITC)، إن الهدف من هذا القسم من الفصل هو وصف الدافع الذي جعل ITC تقوم بإعداد ذلك الدليل، لتوفير بعض المعلومات عن خلفية إعداده وأخيراً وصف الخطوط الرئيسة الاثنتين والعشرين وعرض أسباب استخدام كل واحد منها على حدة.

في ظل الحقيقة أن "المراهنة القوية" ترافق غالباً نتائج الدراسات التعليمية للأداء عبر الثقافات أو المقارنات الدولية (انظر، الاهتمام على مستوى عال اليوم لدعم دراسات الأداء المقارنة الدولية مادياً). فإن الحاجة إلى دليل عملي جيد طور من قبل خبراء لاستخدامه في تكييف الاختبارات وتأسيس نظام معادلة الدرجات النهائية بدت واضحة للهيئة الدولية للاختبارات (ITC) منذ 1992. كانت المقاييس التقنية أو الخطوط الرئيسة لتقويم ممارسة تطوير الاختبارات، جدارة التقويم، إن صدق التقويم متوفرة في بلدان كثيرة (انظر، Aera, APA, & NCME, 1985, 1999)، ولكن لم يكن هناك أي اهتمام لإعداد دليل لتكييف الاختبارات وتأسيس نظام معادلة الدرجات النهائية. على سبيل المثال في مقاييس اختبارات (AERA, APA & NCME) التي نشرت في 1985 (التي كانت من أهم مقاييس الاختبارات في الولايات المتحدة حتى نشرت معايير الاختبارات عام 1999) ثلاثة مقاييس فقط تناولت موضوع تكييف الاختبارات بشكل مباشر. في كندا، بلد ثنائي اللغة، ثلاثة معايير فقط تناولت تكييف الاختبارات في "مقاييس الاختبار في الجمعية النفسية الكندية (كانت المقاييس متوفرة عام 1993).

تعالج الهيئة الدولية للاختبارات ذلك النقص بإعداد مجموعة من الخطوط الرئيسية لتكييف الاختبارات (انظر هامبلتوت، 1994، فان دي فيفر وهامبلتون، 1996)، والذي يشار إليه "دليل الهيئة الدولية للاختبارات لتكييف الاختبارات"، يحدد الجدول 1.1 هوية الهيئات الثمانية التي ساهمت في إعداد الدليل. الجدول 2 يحدد هوية أعضاء اللجنة التي قامت بالعمل لمدة ثلاث سنوات لإعداده.

إن دليل تكييف الاختيار منظم في أربعة أقسام: المحتوى، تطوير وتكييف الاختبار، إدارة الاختبار وتفسير وتوثيق الدرجات النهائية.

كان رأي اللجنة التي أعدت الدليل أنه سيكون مريحاً أكثر في الاستخدام إذا نظم في فئات ذات هدف أساسي. تناولت الخطوات الرئيسية في منشأ المحتوى

تكافؤ المفهوم في لغة المجموعة المستخدمة للاختبار. تتضمن فئة تطوير وتكييف الاختبار موضوعات تظهر في عملية التكييف، بداية باختيار المترجمين إلى طرق الإحصائيات لتحليل المعطيات التجريبية في تقصي تكافؤ الدرجات النهائية. أما الفئة الثالثة، إدارة الاختبار، فإنها تناولت طرق إدارة الاختبار مع مجموعات متعددة اللغات وبداية باختيار الإداريين بين البنود في الاختبار إلى تحديد مدة الاختبار. الفئة الرابعة تختص بتفسير وتوثيق الدرجات النهاية. كالعادة، أعد الباحثون توثيقاً قليلاً جداً عن عملية التكييف لإثبات صدق الاختبار المكيف وكانت الأخطاء في تفسير الدرجات النهائية للاختبارات المتعددة اللغات شائعة جداً. إن دليل ITC لتكييف الاختبارات يعالج كل الأمور في ذلك المجال.

الجدول ١,١

الجمعيات المشاركة في إعداد الدليل العالمي لتكييف الاختبارات:

الهيئة الدولية للاختبار	(ITC)
الجمعية الأوروبية للتقويم النفسي	(EAPA)
مجموعة ناشري الاختبارات الأوروبية	(ETPG)
الجمعية الدولية النفسية عبر الثقافات	(IACCP)
الجمعية الدولية لعلم النفس التطبيقي	(IAAP)
الجمعية الدولية لتقويم الأداء التربوي	(IAA)
الجمعية الدولية للاختبارات اللغوية	(ILTA)
الاتحاد الدولي للعلوم النفسية	(IuPsyS)

الجدول ٢,١

أعضاء اللجنة والمنظمات التي يمثلونها:

رئيس اللجنة

روناد ك. هامبلتون	(ITC)

جامعة مستشوست، أمهرست، الولايات المتحدة

أعضاء اللجنة:

جلين بدجل (ITC)

جمعية الممرضات الكندية، كندا

روب فيلثام (ETPG)

نفرنلسن، إنجلترا

روكيو فرناندز بالستيروز (EAPA)

جامعة أوتونوما، اسبانيا

جون هـ. أ. ل دي جونج (ILTA)

سيتو/ هولندا

أنجرن مونك (IEA)

الإحصائيات السويدية/ السويد

جوزيه مونيز (ITC)

جامعة أوفيدو، إسبانيا

يب بورتينجا (IACCP)

جامعة تيلبيرغ، هولندا

اسيك سفاسير (IuPsyS)

جامعة هاستي، تركيا

تشارلز سبيلبرغر (IAAP)

جامعة جنوب فلوريدا، الولايات المتحدة

فون فان دي فيفر (ITC)

جامعة تيلبيرغ، هولندا

جانس من. زعل (ITC)

GITP الدولية، هولندا

زميل باحث

أنبيل كانجي

جامعة مساشوست، امهرست، الولايات المتحدة

وقد أقرت الهيئة الدولية للاختبارات التعريف التالي لدليل تكييف الاختبارات: "إن دليل تكييف الاختبارات هو مزاولة مهنة تعد مهمة لإدارة وتقويم التكييف أو تطور مواز للاختبارات النفسية والتربوية للاستخدام في مجتمعات مختلفة". إن الخطوط الرئيسة المقدمة من الهيئة الدولية للاختبارات يمكن تلخيصها في النقاش التالي في جدول 3.1 (طبعت كمسودة من قبل في هامبلتون، 1994 وفان دي فيفر وهامبلتون، 1996). تلك الخطوط العريضة موجودة في هذا الفصل مع بعض التعديلات البسيطة في التقرير الأخير للجنة (ITC, 2001) تم وصف كل خط من تلك الخطوط (أ) الأسباب المنطقية لحصر تلك الخطوط، (ب) الخطوات لتطبيق تلك الخطوط، (ج) لائحة الأخطاء الشائعة، (د) ومجموعة من المراجع. هناك نموذج كامل لأحد تلك الخطوط في الجدول (4.1) ويليه وصف مقتضب لكل من الخطوط والأسباب المنطقية لوجودها ضمن اللائحة.

المضمون:

1- ج. أ: يجب تقليل تأثير الاختلافات الثقافية غير الضرورية في أسباب الدراسة الأساسية إلى الحد الأدنى.

الأسباب/ الشرح: هنالك الكثير من العوامل المؤثرة في المقارنة عبر اللغات/ الثقافات التي يجب أخذها بعين الاعتبار عند مقارنة مجموعتين أو أكثر من ذوي خلفية لغوية/ ثقافية مختلفة، خاصة عند تطوير اختبار أو تكييفه أو عند تفسير الدرجات النهائية. على كل حال، من الضروري أن لا يتم التفكير بهذه العوامل فقط بل يجب القيام بخطوات عملية، إما بالإقلال منها أو حذف تأثيرات العوامل غير المرغوب بها في أي مقارنة عبر اللغات/ الثقافات. على سبيل المثال المستويات المختلفة لدوافع المشاركين في الاختبار في بحث حديث للتقويم الدولي للتطور التربوي هو أحد الأسباب لاختلاف أداء المشاركين في الاختبار في تلك الدول (وينر، 1993).

2- ج. د: يجب تقييم تداخلات التراكيب التي يجري تقديرها في الاختبار ضمن المجموعات التي تجري الاختبار.

الأسباب/ الشرح: لا تتوقف الاختلافات الموجودة بين الثقافات واللغات المختلفة للمجموعات على اختلافات تقاليد، قواعد السلوك والقيم ولكن على رؤية العالم وترجماتها، وبذلك من الممكن أن يفسر التركيب ذاته ويفهم بطرق مختلفة كلياً في ثقافتين مختلفين. على سبيل المثال: إن مفهوم "الذكاء" موجود في كل الثقافات تقريباً ولكن في الثقافات الغربية يرتبط هذا المفهوم مع التقديم السريع للإجابات بينما يرتبط في الثقافات الشرقية مع التفكير، الاستجابة، وقول الشيء الصحيح (لونر، 1995)، يجب على الباحثين التأكد من أن التركيب الذي يقاس في اختبار مجموعة من ثقافة/ لغة المصدر الأساسي يمكن أن يوجد بذات الشكل والتواتر في الثقافات الأخرى التي تجري دراستها.

جدول (3-1)

دليل الهيئة الدولية للاختبار ITC لتكييف الاختبار:

ج. 1 (1) يجب تخفيض تأثيرات الاختلافات الثقافية التي ليس لها أهمية للهدف الأساسي للدراسة إلى أقل حد ممكن.

ج. 2 (2) يجب تقويم مقدار التشابك في البنية المقاسة في اختبار المجموعات المطلوبة.

تطور الاختبار وتكيفه

د. 1 (3) يجب على الذين يقومون بعملية التطوير والناشرين التأكد من أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية الثقافية للمجموعات المقصودة.

د . 2 (4) يجب على الذين يقومون بعملية التطوير والناشرين إقامة الأدلة بأن اللغة المستخدمة في تعليمات الاختبار، إرشادات الدرجات، وفي البنود مناسبة للغة وثقافة جميع المجموعات التي ستقوم بالاختبار.

د . 3 (5) يجب على المطورين/ الناشرين إقامة الدليل على أن اختيار أسلوب الاختبار، هيكلية البنود، قواعد الاختبار وإجراءات أخرى مألوفة للمجموعات المقصودة.

د . 4 (6) يجب على المطورين/ الناشرين إقامة الدليل على أن محتوى البنود والمواد الأخرى (المنبهة) مألوفة للمجموعات المقصودة.

د . 5 (7) يجب على المطورين/ الناشرين جمع دليل النقد العقلاني، اللغوي والنفسي، لتحسين دقة عملية التطوير وجمع الدليل على تكافؤ كل النسخ في اللغات المختلفة.

د . 6 (8) يجب على المطورين/ الناشرين التأكد من أن خطة جمع المعطيات تسمح باستخدام أساليب إحصائية مناسبة لإقامة تكافؤ البند والبنية في نسخ الاختبار في اللغات المختلفة.

د . 7 (9) يجب على المطورين/ الناشرين استخدام أساليب إحصائية مناسبة كي يستطيعوا (أ) إقامة التكافؤ في لغة النسخ المختلفة للاختبار.

(ب) التعرف على العناصر التي يمكن أن تحدث مشكلات أو تكون غير مناسبة لإحدى المجموعات المشاركة.

د . 8 (10) يجب على المطورين/ الناشرين توفير معلومات عن صدق الاختبار المكيف للمجموعة المقصودة.

د . 9 (11) يجب على المطورين/ الناشرين توفير الدليل الإحصائي عن تكافؤ البنود لكل المجموعات المقصودة.

د. 10 (12) يجب عدم ربط بنود الاختبار المكيف غير المتكافئ للمجموعة المقصودة مع الدرجات العامة للمقياس. على كل حال يمكن أن تكون تلك البنود مفيدة لإعطاء تقرير عن درجات كل مجموعة على حدة.

الإدارة

أ. 1 (13) يجب أن تكون أوجه المحيط/ البيئة التي تؤثر في إدارة الاختبار متشابهة إلى أقصى حد عبر المجموعات التي تجري الاختبار.

أ. 2 (14) يجب على مطوري الاختبار والإداريين محاولة توقع المشكلات التي يمكن حدوثها واتخاذ الإجراءات المناسبة لمعالجة كل المشكلات وذلك بإعداد مواد وإرشادات مناسبة.

أ. 3 (15) يجب على الإداريين أن يدركوا العناصر المتعلقة بالمواد المحفزة، الإجراءات الإدارية وطرق الاستجابة التي قد تخفض صدق الاستنتاجات التي تم الحصول عليها من الدرجات.

أ. 4 (16) يجب أن تكون لغة إرشادات الإداريين في كلتا اللغتين، لغة المصدر واللغة المستخدمة في الاختبار مقللة من المتغيرات غير المرغوب بها عبر المجموعات.

أ. 5 (17) يجب أن يعطي كتيب الاختبار وصفاً دقيقاً لكل أوجه الاختبار وطريقة إدارته التي تتطلب الدقة في تطبيق الاختبار في محيط ثقافي جديد.

أ. 6 (18) يجب أن لا يكون الإداريون فضوليين، ويجب أن تكون العلاقة بين الإداريين والذين يجرون الاختبار قليلة إلى أدنى حد. يجب اتباع القواعد الواضحة التي جرى وصفها في كتيب الاختبار.

التوثيق/ تفسير المقياس

1 . 1 (19) عندما يكيف الاختبار للاستخدام في مجموعة أخرى، يجب توثيق التغيرات مع الدليل الذي يدعم تكافؤ النسخة المكيفة للاختبار.

1 . 2 (20) يجب أن لا تؤخذ اختلافات الدرجات لنماذج المجموعات التي قامت بالاختبار كقيمة ظاهرية. على الباحث مسؤولية إقامة الدليل على معنى الاختلافات من الأدلة التجريبية.

1 . 3 (21) يمكن إقامة المقارنات عبر المجموعات فقط على مستوى الثوابت التي أقيمت للمقياس الذي تنقله الدرجات.

1 . 4 (22) يجب على المطورين توفير معلومات محددة عن الطرق التي يمكن أن تؤثر فيها المفاهيم الاجتماعية/ الثقافية والبيئية للمجموعات على أداء الاختبار، كما يجب عليهم اقتراح إجراءات لتحليل تلك التأثيرات في تفسير النتائج.

الجدول 1- 4

نموذج للخطوط الأساسية د. 1 في شكله العام

الخط الرئيس د. 1: شروط أساسية عامة ومهنية

يجب على المطورين/ الناشرين التأكد أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية والثقافية للمجموعات المقصودة.

الأسباب/ التفسير:

إن خبرة وتجربة المترجمين يمكن أن تكون أكثر الأوجه أهمية في عملية تكييف الاختبار بأجمعها لأنهم يؤثرون بشكل كبير على جدارة وصدق الاختبار. (براكون وبارونا، 1991). على سبيل المثال: عمد المترجمون الذين يتمتعون بالمعرفة التقنية أو تفاصيل الحقل المترجم إلى الترجمة الحرفية التي يمكن أن تشكل فهمًا خاطئًا

للمجموعة المستهدفة وتهدد صدق الاختبار (هامبلتون وكانجي، 1995). بناء على ذلك فإن اختيار مترجمين مؤهلين جيدين عامل مهم في عملية تكييف الاختبار. بالرغم من أن الخبرة في اللغتين شرط أساسي فإن المعرفة والتجربة (أ) للثقافتين، (ب) محتويات الاختبار، (ج) ومبادئ تطوير الاختبار خاصة كتابة البنود، يجب أن تكون أحد الشروط الأساسية في اختيار تدريب المترجمين. هنالك حاجة لوجود فريق عمل للقيام بعملية تكييف دقيقة؛ لأنه من الواضح أنه ليس من المعقول أن يتمتع فرد واحد من المترجمين بكل تلك الشروط الأساسية.

1- التأكد كحد أدنى أن المترجمين مؤهلون وذوو خبرة في كل من اللغتين، لغة المصدر واللغة المستهدفة وثقافتهما (بتشر وغراسيا، 1987). إن الشهادات أو الخبرة السابقة شروط أساسية مهمة. على سبيل المثال: ليس بالإمكان افتراض أن ثنائيي اللغة متمكنون في اللغتين أو مطلعون على كلتا الثقافتين بشكل متساو.

2- إن معرفة موضوع المادة شرط أساسي لأي مترجم يقوم بتكييف اختبارات إذا لم تكن له معرفة بموضوع المحتوى على الأقل فإن دقة موضوع المحتوى قد تفقد. يجب أن يكون للمترجمين الذين ليس لهم اطلاع على معرفة معينة في حقل الترجمة الاطلاع المسبق على موضوع المحتوى كجزء من عملية التكييف.

أين يمكن أن يعيش طائر له أقدام كفيره (كأقدام البط)؟

أ - في الجبال.

ب- في الغابات.

ج - في البحر.

د - في الصحراء.

عندما تُرجم هذا السؤال من الإنجليزية إلى اللغة السويدية، أصبحت "الأقدام الكفين" "أقدام سابحة" وبذلك أصبحت الإجابة واضحة للطلاب السويديين عن

مكان عيش الطائر. إن مترجماً ذا معرفة بقواعد كتابة البنود كان سيلاحظ دون شك الخطأ في الترجمة ويقوم بتعديلها.

4- من المفضل أن يقوم فريق من الاختصاصيين بمشروع تكييف الاختبار (انظر، غريسي، 2003). يجب أن يشارك المترجمون في فريق المشروع وأن يكون لهم دور في تقرير عملية الترجمة، وأن تطلب آراؤهم وأفكارهم وتقبل. إن تلك الطريقة حسب "برسلين" (1986) تستطيع تحسين نوعية التكييف. إن طريقة فريق عمل تساعد في (1) تمكن استخدام طريقة الترجمة الراجعة (انظر الخطة 5 في الأسفل)، (2) تسمح للمترجمين بمقارنة ومناقشة أعمالهم وبذلك تُحسن الصلة والنوعية للترجمة، (3) تساعد في التأكد من أن المعرفة المختصة في كل الحقول المطلوبة يمكن الوصول إليها.

5- إن استخدام فريق عمل من المترجمين الذين يعملون إما منفردين أو ضمن مجموعات صغيرة لتكييف الاختبار هو خطه عمل ممكنة. من الممكن لاحقاً القيام بمقارنة التقويم الفردي للاختبار وحل الاختلافات لتقديم الترجمة الأفضل، هنالك إجراء آخر وهو استخدام مطورين ومترجمين أحاديي اللغة في وقت واحد، يقوم المترجمون بترجمة/ تكييف الاختبار ثم يقوم المطورون أحاديو اللغة بتحرير الاختبار في اللغة المستهدفة، ومن ثم يقوم مترجم/ مطور ثنائي اللغة بتقويمه (برسلين، 1986). لاحظ "برسلين" (1986) أن ميزة تلك الطريقة هي أن مطوري الاختبار أحاديي اللغة يستطيعون إعادة كتابة الاختبار ويجعلونه أكثر وضوحاً واستحساناً لطلاب اللغة المستهدفة، كما أن تلك الخطة تقلل من المواقف، حيث تكون النسخة الجديدة ضعيفة، ولكن من الممكن إغفال تلك المشكلة لأن وجود مترجم عالي الخبرة يستطيع تقديم نسخة ترجمة راجعة ممتازة من نسخة اختبار سيئة في اللغة المستهدفة. أما في حالة وجود مترجم واحد فقط فمن المفضل استخدام مترجم من مجموعة اللغة المستهدفة. في

ذلك الموقف يستطيع المترجم على الأقل مناقشة نسخة اللغة المستهدفة مع شخص آخر من مجموعة اللغة المستهدفة الذي يستطيع الإشارة إلى مواطن المشكلة وربما يقترح التنقيح أيضاً.

أخطاء شائعة:

1- اختيار المترجمين أو الأشخاص من معارف مطور الاختبار (أصدقاء، جيران) لأنهم ثنائيو اللغة قد أثبت أنه اختيار غير ناجح (برسلين، 1986).

2- الفشل في التأكد من اختيار المترجمين ذوي الاطلاع على محتوى الاختبار والذين لهم خبرة في تطوير الاختبار، لقد تم الإبلاغ عن حدوث تلك الأخطاء في كثير من الدول.

3- لم يعط المترجمون الوقت الكافي للقيام بأعمالهم، وقد تم الإبلاغ عن هذه الأخطاء أيضاً.

تطوير الاختبارات وتكييفها:

1. د. 1: يجب على الذين يقومون بتطوير الاختبارات ونشرها التأكد أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية الثقافية للمجموعات التي تجري الاختبار.

الأسباب/ الشرح إن أسباب تلك الخطة مع الأجزاء الأخرى لتوصيف هذا الدليل تظهر في الجدول (4.1) الذي يستخدم كنموذج للمعلومات المتوفرة عن كل خطة في التقرير النهائي (انظر ITC، 2001).

2. د. 2: يجب على القائمين على تطوير الاختبار ونشره توفير الدليل على أن اللغة المستعملة في إرشادات الاختبار، قواعد الدرجات النهائية والبنود الموجودة في الاختبار كلها مناسبة لجميع الثقافات ولغات المجموعات التي يستهدفها الاختبار.

الأسباب/ الشرح: إن أحد أسباب سوء تكييف اختبار الدراسة عبر الثقافات هو وجود خطأ في نسخة الاختبار في لغة المصدر؛ وذلك يسبب صعوبة في التكييف. وهناك سبب آخر هو أنه من الممكن أن تكون المفاهيم، والتعابير والأفكار المستخدمة في لغة اختبار المصدر ليس لها مرادف في اللغة المستهدفة. إن أحد الأسباب الكثيرة لنجاح الدراسات الحديثة التي قام بها TIMSSو OECD / PISA هو الجهد الفعلي الذي بذل في تطوير الاختبار في لغة المصدر مع وضوح التنظيم ووجود مواصفات الاختبار وبعض بنود للتطوير والاختبار الميداني، وبعض الفعاليات المرافقة لتطوير اختبار مناسب.

من الضروري أيضاً التأكد أن المفردات المستعملة في اختبار متعدد اللغات متشابه من حيث درجة صعوبة الكلمات، نصوص القراءة، استعمال القواعد، أسلوب الكتابة، التنقيط، يجب الحذر عند استعمال الاختبار لتقويم مقدرة المشاركين في الكتابة والقراءة (الصغار والبالغين).

3. د. 3: يجب على المطورين والناشرين توفير الدليل على أن اختيار الأسلوب، الشكل العام، وكل الخطوات المتبعة في الاختبار مألوفة للمجموعة المستهدفة.

الأسباب/ الشرح بعض الأشكال العامة (أسئلة عديدة الإجابات، المقالة، 5 درجات مقياس التقدير) والاصطلاحات والإجراءات في إعطاء الإرشادات الاختبارية وفي تقديم بنود الاختبار قد لا تكون مألوفة في كل المجتمعات. هنالك اختلاف في الاستعمال اللغوي في إرشادات الاختبار، التنسيق واستعمال الرسوم البيانية، طريقة التقديم (الأقلام، الأوراق، الكمبيوتر). لتحقيق المساواة من الضروري أن يكون كل ما سبق مألوفاً للمجتمعات التي يُجري تكييف الاختبار لها. وهذا يتضمن تطوير مواد علمية مكثفة لتقليص التحيز الناتج عن عدم الألفة في بعض وجوه عملية التقويم.

4. د. 4: يجب على المطورين والناشرين إثبات أن كل بنود المحتوى والمادة المحفزة مألوفة لدى المجموعات المستهدفة.

الأسباب/ الشرح. إذا ثبت أن أي اختبار مكيف هو أسهل أو أصعب قراءة أو فهماً بسبب بعض البنود في المحتوى فإن هذا سيكون مصدر تحيز آخر. في بعض الدول في العالم تستعمل وحدات مختلفة للمقادير، على سبيل المثال الوزن، الطول والنقود. يمكن أن يكون تكييف الاختبار أكثر صعوبة للمجموعات المستهدفة إذا كانت الوحدات المستخدمة غير مألوفة أو إذا كانت هنالك بعض العمليات الحسابية (انظر هامبلتون، يون وسليبتر، 1999) وإذا كان هناك مواد محفزة (أشكال، أرقام، جداول، أو علامات حدود) غريبة عليهم.

5- د. 5: يجب على المطورين الناشرين جمع الدلائل العقلانية، اللغوية والنفسية، لتحسين الدقة في عملية التكييف وجمع الدلائل على تساوي كل نسخ الاختبارات في اللغات المختلفة.

الأسباب/ الشرح. يجب تقييم الأسئلة، التمارين، تقدير المقاييس المرادفة في اللغات والثقافات المختلفة. إن الطرق العقلانية في القيام بالترجمة المترادفة تقوم على قرارات المترجمين أو مجموعات المترجمين. إن الطريقتين المستخدمتين في الترجمة: الخطة المقدمة والخطة الراجعة، اللتين جرى ذكرهما في بداية هذا الفصل، فيهما بعض الأخطاء ولذلك من الصعب جداً أن توفر الخطط العقلانية الدلائل الكافية لصدق الاختبار المكيف.

6. د. 6: على المطورين/ الناشرين التأكد أن خطة جمع المعطيات تسمح باستخدام تقنيات إحصاء مناسبة لإقامة التساوي في الشكل والمحتوى في نسخ الاختبار في لغات مختلفة.

الأسباب/ الشرح. إن خطة جمع المعطيات تعود إلى الطريقة التي جمعت بها تلك المعطيات كي يجري التساوي في نسخ الاختبارات المكيفة. إن أول شروط جمع المعطيات هو أن النماذج يجب أن تكون كثيرة بشكل كبير كي يكون هناك إمكانية الحصول على معلومات إحصائية متوازنة. مع أن هذا الشرط ضروري لأي نوع من الأبحاث إلا أنه مهم بشكل خاص في تقدير صدق الاختبار المكيف لأن الأعداد الكبيرة الكافية للنماذج قد يكون لها دور في جمع المعطيات اللازمة لإثبات التعادل في الاختبار.

إن الخطة في الدراسة التجريبية هي مجموعة وظائف (أ) طبيعة المشاركين (أحادي أو ثنائي اللغة)، (ب) نسخة الاختبار المستخدمة (الأصلية، المكيفة، المكيفة الراجعة) (ج) خطة جمع المعطيات المحددة (تناقش بشكل مفصل في د. 7). قدم سيرغي (1997) نقاشاً عن المعضلات والموضوعات في ربط اختبارات متعددة اللغات مع مقياس عام. قدم ودكوك ومونوز ستاندوفل (1993) نموذج ربط مقياس اختبار مع اختبار عبر اللغات مستخدماً IRT انظر في فصل لاحق في هذا الكتاب).

7. د. 7: على المطورين/ الناشرين تطبيق تقنيات إحصائية مناسبة في (أ) إقامة التساوي في لغة الاختبار المستخدم (ب) تعيين المعضلات أو العناصر التي قد تكون غير مناسبة للاستخدام ضمن إحدى المجموعات.

الأسباب/ الشرح. تقدم التقنيات الإحصائية معلومات مفيدة لتقويم تساوي الاختبارات المطورة في أكثر من لغة (فان دي فيفر ولينونغ، 1997، 2000، فان دي فيفر ونانزر، 1997، انظر فصل لاحق). يجب أن تستخدم تلك التقنيات في توفير إضافات إلى تقنيات المقارنة لكونها قادرة على تعيين بنود الاختبار غير المتوافقة التي لم يتم اكتشافها عند استعمال تقنيات المقارنة. هناك ميزة أخرى وهي أن التقنيات الاحصائية تستخلص

المعلومات مباشرة من المشاركين في الاختبار، من مضمون إدارة الاختبار العقلية وبذلك تكون تلك التقنيات مفيدة جداً في تعيين البنود التي قد تشكل بعض الصعوبات في التطبيق.

8. د. 8: يجب على المطورين/ الناشرين توفير المعلومات عن صدق الاختبار المكيف ضمن المجموعة المستهدفة.

الأسباب/ الشرح. إن الاختبارات الموجودة يجري تطويرها وتوحيدها غالباً للاستخدام في ثقافة واحدة وتُكيف للاستخدام في ثقافة أخرى. يمكن توفير الوقت والنفقات إذا جرى تكييف الاختبارات الموجودة (برسلين، 1986). على كل حال فإن كثيراً من التراكيب تكون غير مفهومة دون بعض التعديلات الأساسية للاستخدام في الثقافات الأخرى. طرحت عدة نماذج في هذا الفصل، الذكاء، نوعية الحياة اليومية، والإنجازات الرياضية. في بعض الأحيان من الممكن أن يكون الاختبار غير جدير بالترجمة وبذلك يمكن توفير الوقت، الجهد والمال، حتى في حال وجود ذلك التركيب في اللغة أو الثقافة الثانية من الممكن وجود تفاوت في المظاهر السلوكية والتفسيرات بشكل واضح (لونر، 1990). يجب جمع الأدلة على صدق التركيب في كل مجموعة تجري الاختبار. كما هو معروف فإن استقصاء صدق التركيب يستغرق وقتاً للتخطيط والتطبيق لكونه شاملاً ويتضمن دراسات ومنهجيات متنوعة منها اختبارات متداخلة، متعلقة بالمقياس، تجريبية، متعددة السمات والطرق (انظر. فان دي فيفر وتانزر، 1997).

9. د. 9: على المطورين/ الناشرين توفير إثبات إحصائي عن تساوي البنود في كل المجموعات المستهدفة.

الأسباب/ الشرح. إن أحد أهم التحليلات الإحصائية في صدق اختبار للاستخدام في لغة/ ثقافة واحدة أو أكثر في دراسة بنود متحيزة أو كما يشار إليها حالياً "دراسة تفاوت أداء البند DIF" (هولند ووينر، 1993، سيرغي وآلالوف، 2003، وعدة فصول في هذا الكتاب). يتطلب مساندة تكافؤ اختبار لمجموعتين ثقافتين مختلفتين وجود إثبات على أن أعضاء المجموعتين لهما الكفاءة الواحدة، يجب عليهم الأداء المتكافئ في كل بند. عندما لا يكون الأداء متساوياً يجب أن يكون هناك سبب وجيه أو يلغى البند من الاختبار. هذا لا يعني عدم وجود اختلاف أداء عام في الاختبار. بشكل عام من المتوقع وجود اختلافات. هذا يعني عندما تجري مقارنة المجموعتين في التراكيب المقاسة في الاختبار، في حال وجود الاختلافات، عندئذ تكون دراسة تفاوت DIF أداء البند موجودة ويجب دراسة خواص البند بحذر قبل استخدامه في أي اختبار. إن البنود المشار إليها "DIF" قد تكون مسببة للمشكلات بسبب سوء الترجمة أو بسبب استعمال مصطلح، موقف، أو تعابير غير معروفة أو مألوفة إلى إحدى المجموعات الثقافية. هناك أسباب أخرى أيضاً ربما المهارة التي يجري اختبارها في تلك البنود ليست جزءاً من ذخيرة ثقافة مجموعة اللغة المستهدفة أو ربما يكون شكل البند غير مألوف، إن تقرير أسباب الاختلاف مهم لأنه يوثر في تقرير ما يجب القيام به بشأن ذلك البند.

يكون لتقديم الخطوط الرئيسة معنى عندما يكون هناك إثبات أن التركيب له صلة مع المجموعة الحضارية المستهدفة، وأن هناك إثباتاً أنه قد جرى التحقق من الترجمة والتكييف بحذر (ربما بواسطة خطة الترجمة المقدمة). هناك ثلاث منهجيات يمكن استخدامها بشكل أساسي لإجراء عدة نماذج للتحليلات المطلوبة في تلك الخطوط الرئيسة:

(أ) إجراءات (IRT انظر اليس، 1989، 1999، اليس وكيميل، 1992).

(ب) إجراء مانتل - هانزل (MH) وما يتبعه (انظر، هامبلتون، كلوس، يزور، وجونز، 1993، هولند وثابر، 1988، هولند ويونر، 1993، سيرغي وآلالوف، 203).

(ج) إجراء منطقي ارتدادي (LR) سوامينثان وروجرز، 1990) إن كل هذه المنهجيات شرطية بمعنى أن المقارنة تجري بين مجموعتي أشخاص مثلاً: (إنجليز - فرنسيين) الذي جرى الافتراض أنهم متماثلون في المهارات المقاسة في الاختبار. في إجراء IRT يُطابق الممتحنون باستخدام درجات المهارة المقدرة (تقدر باستخدام نموذج درجات البنود). كل تلك الإجراءات مجتمعة تستخدم الدرجات النهائية للاختبار لمقارنة الممتحنين. وكل تلك المنهجيات تعطي نتائج ثابتة وصادقة في حال كانت النماذج كثيرة ومتوفرة وأنه قد تم إنجاز الإجراء بشكل صحيح وأن النتائج قد فسرت بعناية.

يحتاج إجراء LR و MH إلى عينة حجمها 200 من كل مجموعة ثقافية على الأقل. بشكل عام تحتاج إجراءات IRT حقيقة إلى عينات أكبر.

10. د 10: يجب عدم استخدام البنود غير المتكافئة للمجموعة الثقافية المستهدفة في ربط الاختبار المكيف مع مقياس الدرجات النهائية العامة المقدمة. على كل حال تلك البنود قد تكون مفيدة لتقديم الدرجات في كل مجموعة بشكل منفصل.

الأسباب/ الشرح. قد تعتبر بنود في الاختبارات المكيفة بعض الأحيان غير متكافئة بسبب التكييف السيئ أو كونه غير مناسب في تلك الثقافة (هلن، 1987).

لا يمكن استخدام تلك البنود في ربط النسخة المكيفة للاختبار مع مقياس الدرجات النهائية العامة لأنها توفر معلومات مختلفة عن المجموعات التي جرت مقارنتها. على كل حال تقدم البنود المكيفة بشكل جيد والتي جرى تعريفها بأنها غير متكافئة (غير مناسبة ثقافياً) معلومات مفيدة عن تلك الثقافات والحضارات. إن معرفة مصدر عدم التكافؤ لتلك البنود قد يقدم معرفة أدق عن ثقافة ولغة تلك المجموعة وهذا يؤدي إلى زيادة فهم تلك المجموعة (اليس، 1991).

الإدارة:

1. أ. 1: إن ظروف البيئة التي تؤثر في إدارة الاختبار يجب أن تكون متشابهة قدر الإمكان عبر المجموعات الثقافية التي يستهدفها الاختبار.

الأسباب/ الشرح. يختلف عدد الصعوبات المتوقعة في إدارة الاختبار باختلاف الفروق اللغوية والثقافية بين المجموعات التي يجري اختبارها أو بين ثقافة المجموعة التي جرى اختبارها أولاً وثقافة المجموعة التي ستقوم بالاختبار. هناك حاجة لمعرفة ثقافة ولغة المجموعة المستهدفة لكي تستطيع العمل في هذا الدليل. من المتوقع أن يواجه المطور بصراحة الصعوبات التي تؤثر في عملية المقارنة وأن يدرس الإجراءات الضرورية، يجب تقديم الدلائل التجريبية لدعم مطالبة المقارنة. إذا لم يكن ذلك ممكناً يمكن تقديم مناقشة عقلانية لتبرير استخدام الاختبار المكيف عبر الثقافات.

2. أ. 2: يجب على المطورين والإداريين محاولة توقع أنواع الصعوبات واتخاذ الإجراءات المناسبة لمعالجتها وذلك بإعداد إرشادات ومواد مناسبة.

الأسباب/ الشرح. يجب أن يكون لدى مطوري الاختبار معلومات جيدة عن تطور الاختبار عبر الثقافات، إضافة إلى ذلك يجب أن يتمتعوا بالخبرة الكافية كي يستطيعوا إدارك تعقيدات وخاصية إدارة الاختبار عبر الثقافات. إحدى الطرق العلمية هي إعداد جدول عن الصعوبات التي تحدث غالباً وقد تهدد صدق الاختبار.

إن معرفة اللغة والثقافة الدقيقة للمجموعة المستهدفة ضروري جداً؛ على سبيل المثال فإن درجة 3 أو 4 في مقياس الدرجات في تركيا هو الأفضل، وتشكل درجات أعلى مشكلات في الدلالات اللغوية.

3. أ . 3: يجب أن يكون لدى الإداريين إحساس دقيق بعدد من العوامل المتعلقة بالمواد المحفزة، الإجراءات الإدارية، وأساليب الاستجابات التي قد تؤثر على صدق الاستنتاجات التي تم الحصول عليها من الدرجات النهائية.

الأسباب/ الشرح. قد تكون شروط إدارة الاختبار مصدر متغيرات غير مقصودة في الدرجات النهائية. كي نستطيع مضاعفة صدق وعملية مقارنة درجات الاختبار عبر مجموعات ثانية، يجب وصف الأسباب التي قد تؤدي إلى اختلافات في الدرجات النهائية.

4. أ . 4: يجب وجود التعليمات الإدارية في لغة المصدر وفي اللغة المستهدفة لتقليص تأثير أسباب الاختلافات (غير المرغوب فيها) عبر المجموعات المختلفة.

الأسباب/ الشرح. تخاطب الدراسات عبر الثقافات غالباً مجموعات ذوي خلفية مختلفة جداً. عندما يبدأ الطلاب المشاركون في الاختبار بالإجابة عن الأسئلة/ التمارين يجب تقليص تأثير مصادر الاختلافات غير المرغوب فيها بقدر المستطاع. إحدى الطرق للقيام بهذا هو إرشادات الاختبار الواضحة.

5. أ . 5: يجب أن يحتوي كتيب الاختبار على كل تفاصيل الاختبار وإدارته التي تحتاج إلى تدقيق في تطبيق الاختبار في محيط ثقافي جديد.

الأسباب/ الشرح. كثير من السمات المتعلقة بإدارة الاختبار ضمن مجموعة لغوية أخرى يمكن أن يتوقعها الذين يقومون بتطوير الاختبار: لذلك يجب على المطورين جمع معلومات عن موضوعات معينة والتي من المحتمل أن تكون متعلقة بالاختبار المكيف في أثناء عملية تطوير صدق الاختبار. في بعض الحالات يحصل المكيف على معطيات من الأقليات الثقافية أو التطبيقات عبر الثقافات الموجودة، يجب وجود المعلومات المتعلقة بإدارة اختبارات تلك المجموعات في كتيب الاختبار.

6. أ . 6 يجب أن يكون الإداري فضولياً كما يجب تقليل العلاقة بين الإداريين والطلاب. تناقش قواعد واضحة عن إدارة الاختبار لاحقاً في الكتيب.

الأسباب/ الشرح. من الممكن أن يكون تأثير الإداري على نتائج الاختبار حقيقياً. إن الهدف هو تقليص ذلك التأثير وذلك بأن يتعهد الإداريون باتباع الإرشادات والإجراءات المعتمدة في إجراء الاختبار؛ من ناحية ثانية قد يكون للإداريين تأثير غير واضح، وغير مرغوب. إن صفات الإداري مثل الجنس، العمر، العرق وحتى طريقة اللباس وأشياء أخرى يمكن أن تؤثر في نتائج الاختبار خاصة إذا كان هنالك إداري واحد فقط. إذا جرى استخدام اختبار جديد مكيف في مجموعة ثقافية ما يمكن أن يكون من الأسهل نسبياً، إذا كان الإداري ينتمي إلى ذات المجموعة، أن نستطيع تحديد صفات الإداريين التي يمكن أن تهدد صدق حصيلة نتائج الاختبار. عندئذ يمكن القيام ببعض الخطوات (كدراسة تجريبية) في حالة عدم تماثل الخلفية الثقافية للإداريين والطلاب الممتحنين خاصة، يجب التحقق من التأثير السلبي المحتمل للإداري واتخاذ الخطوات لتقليل المشكلات المحتملة.

التوثيق/ تفسير الدرجات:

1 . 1 . 1 عندما يكيف الاختبار لاستخدامه ضمن مجموعة مختلفة، يجب توثيق التغيرات في الاختبار مع البراهين الداعمة لتكافؤ النسخة المكيفة للاختبار.

الأسباب/ الشرح. من الممكن أن توفر معلومات عن تفاصيل في تكييف اختبار فكرة ثاقبة إذا كان من المناسب استخدام الاختبار ضمن بيئة محددة. على سبيل المثال: معرفة أن عوامل ثقافية، اجتماعية في ثقافة ما قد تم أخذها بعين الاعتبار - أثناء عملية تكييف اختبار لمتكلمي اللغة الإسبانية في أميركا الجنوبية يمكن أن يكون ذا قيمة عند تقرير عما إذا كان الاختبار مناسباً

لاستخدام المتكلمين بالإسبانية في الولايات المتحدة. يجب توثيق الإجراءات المتبعة في تكييف الاختبار بكاملها في كتيب الاختبار لتسهيل تقويم الاختبار من قبل مستخدمين محتملين. يجب أن يتضمن التوثيق بيان خطوات مفصلة عن الإجراءات بكاملها بما فيها المحاكمة العقلية المستخدمة، الطرق المستخدمة في تقويم البنود وتكافؤ الاختبار المكيف ونتائجه. تفاصيل عن اختيار المترجمين واستخدامهم، أسباب وتبرير استخدام وإدخال بعض البنود ومعلومات عن البنود التي تم تعديلها أو استبعادها، بعض المشكلات الرئيسة التي واجهت سير عملية التكييف وكيف أمكن حلها، كل النواحي المتعلقة بإدارة الاختبار بما فيها اختبار وتدريب الإداريين وتفسير النتائج.

2. 1. 2: يجب أن لا يكون لاختلاف درجات اختيار عينات المجموعات التي أجرت الاختبار قيمة ظاهرية. تقع على الباحث مسؤولية إثبات معنى الاختلاف بدلائل تجريبية (إمبريقية).

الأسباب/ الشرح. يبدو أن الأخطاء العامة في التطبيق هي التي تعطي أهمية محدودة العملية تكييف الاختبار، والتي تفسر اختلافات الدرجات لمجموعة كأنها انعكاس اختلافات حقيقية للتركيب الذي يجري قياسه بواسطة الاختبار. إن تجاهل مشكلات تكييف الاختبار التي تحصل بشكل دوري في التطبيق والحاجة إلى تأكيد صدق الاختبارات في الثقافات التي تجريها قد شوهت صحة نتائج الدراسات الكثيرة عبر الثقافات، إن وجود عملية تكييف موثوقة أساسية لإثبات صدق الاختبار المكيف. في الوقت ذاته حتى في وجود اختبارات مكيفة ممتازة يجب على الباحثين القيام بجهود لتفسير نتائجهم. بشرط معرفتهم الكاملة للثقافات المستخدمة للاختبار. هذا يعني، على سبيل المثال أنه يجب جمع دلائل ثابتة كلما كان ذلك

ممكناً، وإذا لم يكن ذلك ممكناً يجب الحذر التام في تفسير النتائج التي تم الحصول عليها من مجموعات مختلفة.

3. 1. 3: يمكن إجراء المقارنة عبر المجموعات في حالة ثبات المقياس الذي تستند عليه الدرجات.

الأسباب/ الشرح. في بعض الأحيان من الممكن تصنيف درجات اختبار بلغات مختلفة حسب مقياس عام وذلك بغية تسهيل عملية المقارنة للدرجات. عند الحصول على نماذج كثيرة، نماذج إحصائية فعالة مثل إحصائيات IRT (انظر هامبلتون وآخرين، 1991).

يمكن الحصول على معدل مترابط لدرجات اختبار مكيف إذا كان بناء الاختبار متعادلاً في النسخ المختلفة وإذا كانت المعطيات الصحيحة لذلك التعادل متوفرة (انظر د. 6). عند حصول ذلك يمكن القيام بكل أشكال مقارنة الدرجات بما فيها متوسط الدرجات، الانحراف القياسي والتوزيع. غالباً تكون درجات نسخ الاختبار المترجمة إلى لغات مختلفة لم تعدل جيداً وبذلك لا يمكن مقارنة الدرجات مباشرة. ومع ذلك يمكن القيام بمقارنة دور بناء الاختبار في كل مجموعة لغوية. على سبيل المثال في اختبار المهارات المكيف من الإنجليزية إلى الإسبانية، يمكن أن يكون الباحث مهتماً بمقارنة صدق مهارة التنبؤ في الاختبار في كل مجموعة لغوية. إن الغرض الأساسي لهذا الكتيب هو التأكد أن الباحثين لا يقومون بمقارنات غير مبررة لدرجات اختبار نسخ مترجمة للغات متعددة وأنهم يقتصرون في عمليات التفسير على المقارنات التي توفر دلائل صدقها.

4. 1. 4: يجب على مطور اختبار توفير معلومات محددة عن الطرق التي يمكن أن تؤثر فيها الثقافة الاجتماعية، البيئة لمجموعة على الأداء في الاختبار، كما يجب عليه اقتراح إجراءات لبيان أسباب تلك التأثيرات على عملية تفسير النتائج.

الأسباب/ الشرح. إن العوامل المختلفة في أي دراسة عبر ثقافات/ القوميات المتعلقة بأسباب الاختبار يجب اعتبارها عاملاً للحصول على فهمٍ كاملٍ للنتائج (براكون وبراونا، 1991). إن العوامل الاجتماعية/ السياسية المختلفة التي تؤثر على الأداء في الاختبار بشكل ثابت لا تؤخذ بعين الاعتبار (فان دي فيفر وبورتينغا، 1991) غالباً. مثلاً عند مقارنة الأداء الأكاديمي لطلاب من دول نامية وطلاب من دول متطورة، يمكن أن تكون فروق الأداء عائدة إلى عدم إمكان الحصول على مراجع لا عن عدم وجود الإمكانيات أو ربما تكون انعكاساً لنوعية الخدمات التربوية المتوفرة.

الخاتمة

كي يجري تقدير معنى وفائدة أبحاث عبر الثقافات، من الضروري أن يكون الباحثون حذرين في اختيار الإداريين، وأن يستخدموا بنود اختبار مناسبة وأن يسيطروا على عامل السرعة. بالإضافة إلى ذلك فإن المترجمين المتآلفين مع المجموعة المستهدفة ومع ثقافتهم، الذين يعرفون محتوى الاختبار والذين تم تدريبهم على تطوير الاختبار، هم الأشخاص الأكثر مقدرة على تقديم اختبار مكيف صادق. إن اختبار مخطط عقلاني مناسب، مخطط جميع معطيات وتحليل إحصائيات يمكن أن يوفر معطيات قيمة متعلقة ببنود الاختبار وتكافؤ الاختبار عبر مجموعات لغوية وثقافية مختلفة. أما ما يتعلق بتفسير الدرجات، فيجب أن نفكر بحذر بالتفاصيل الخلفية للمتغيرات المؤثرة على الأداء، المناهج المختلفة، مستوى الدوافع والعوامل الاجتماعية/ السياسية التي يمكن أن تكون مهمة بشكل خاص. يجب أن لا تقوم المقارنة بالتركيز على الاختلافات فقط. يمكن أن تؤمن التشابهات بين المجموعات المختلفة اللغات والثقافات معلومات مفيدة ووثيقة الصلة بالبحث.

إن دليل ITC لتكييف الاختبار الذي جرى وصفه في هذا الفصل يؤمن خطوط عمل عريضة للباحثين للقيام بدراسات في تكييف الاختبار ويتوقع أن يكون ذلك

الدليل مع التوصيات المرافقة مفيداً لعدد كبير من الهيئات وأن يحسن نوعية تكييف الاختبارات في العالم وبذلك يساهم في صدق أبحاث عبر اللغات وعبر الثقافات (انظر ITC، 2001). هنالك عدد لا بأس به من المراجع للقراء: قدم كيسنجر (1994). هامبلتون وباتسولا (1999) خطوات مفصلة لمشاريع تكييف الاختبارات، هامبلتون وغيره (1999) تقدموا بنتائج واحد من الاختبارات الأولية الميدانية، تناول هاركنس (1998) موضوعات وطرق مرتبطة بتكييف الاختبار مع التركيز على مقياس التقدير، فان دي فيفر وبورتينغا (1997) قدموا خطة عمل لاستقصاء التهديدات لصدق تفسير درجات اختبار عبر الثقافات.

كما قدم فان دي فيفر وتانزر (1997) سيرغي وآلالوف (2003) لوائح شاملة للإجراءات الاحصائية.

شكر

يظهر هذا الفصل أيضاً في "تقرير البحث التقويمي والقياس السيكولوجي المخبري" رقم 353، جامعة ماستشوست، كلية التربية، امهرست.

يود المؤلف أن يشكر مجلس الكلية لتقديمه الدعم المادي للبحث، على كل حال مجلس الكلية ليس مسؤولاً عن أي خطأ كما أنه لا يجب افتراض مصادقة المجلس على الآراء المقدمة.

يشكر المؤلف فون فان دي فيفر ويب بورتينغا من جامعة تيلبرغ لمساعدتهم في إعداد هذا الفصل كما يشكر يان باتسولا واكتبيل كانجي لمساعدتهم التقنية.

*****

المراجع

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137–164). Newbury Park, CA: Sage.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal, 56*(8), 472–475.

Cziko, G. (1987). Review of the Bilingual Syntax Measure I. In J. C. Alderson & K. J. Krahnke (Eds.), *Reviews of English language proficiency tests.* Washington, DC: Teachers of English to Speakers of Other Languages.

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74,* 912–921.

Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. *Bulletin of the International Test Commission, 18,* 33–51.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177–184.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199–215.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 54–65.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–240.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Method-*

ological advances in cross-national surveys of educational achievement (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission, 18,* 3–32.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1–18.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127–240.

Hambleton, R. K., & Kanjee, A. (1995a). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–160.

Hambleton, R. K., & Kanjee, A. (1995b). Translation of tests and attitude scales. In T. Husen & T. N. Postlewaite (Eds), *International encyclopedia of education* (2nd ed., pp. 6328–6334). Oxford, England: Pergamon.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1–16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15*(3), 270–276.

Harkness, J. (Ed.). (1998). *Cross-cultural equivalence.* Mannheim, Germany: ZUMA.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross cultural psychology. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18,* 115–142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translation. *Journal of Applied Psychology, 67,* 818–825.

International Test Commission. (2001). *International Test Commission guidelines for test adaptation.* London: Author.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics* (Report No. 22-CAEP-01). Princeton, NJ: Educational Testing Service.

Lonner, W. J. (1990). An overview of cross-cultural testing and assessment. In R. W. Brislin (Ed.), *Applied cross-cultural psychology* (Vol. 14, pp. 56–76). Newbury Park, CA: Sage.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43*, 1–14.

Rosansky, E. J. (1979). A review of the Bilingual Syntax Measure. In B. Spolsky (Ed.), *Some major tests: Advances in language testing (Series 1)*. Arlington, VA: Center for Applied Linguistics.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*, 12–19.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–166.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

van Leest, P. F., & Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology*. Amsterdam: Swets & Zeitlinger.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33–51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Boston: Kluwer Academic.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17–24.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29–37.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263–279.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1–21.

Westbury, I. (1992). Comparing American and Japanese achievement: Is the United States really a low achiever? *Educational Researcher, 21*, 18–24.

Woodcock, R. W., & Munoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment, 9*, 233–241.

2

مسائل مفاهيمية ومنهجية في تكييف الاختبارات

فون ج..ر. فان دي فيغر ويب هـ. بورتينغا
جامعة تيلبرغ، هولندا

لنفترض أن عالماً نفسياً قرر عمل نسخة باللغة الألمانية من أحد اختبارات الذكاء الأميركية الذي يختبر المعلومات العامة الذي يتضمن البند التالي: من هو رئيس الولايات المتحدة، قد يكون بند هذا السؤال جيداً لقياس سيكولوجية عينة من الأميركيين. قد يقرر المختص الألماني استخدام ترجمة حرفية لهذا السؤال: من هو رئيس الولايات الألمانية؟ إن السؤال بصيغته الألمانية هو سؤال واضح كمثيله في اللغة الإنجليزية وأن ترجمة راجعة لهذا السؤال ستعطي النص الأصلي الإنجليزي. يضاف إلى ذلك أن الرئيس الأميركي معروف بشكل أفضل من قبل المواطنين الأميركيين، ولكن من المؤكد أنه معروف أيضاً في أوروبا. عند استخدام هذا الاختبار من المرجح أن تكون الإجابة عن هذا السؤال أقل صعوبة في الولايات المتحدة منها في هولندا. وإن هذا الفارق هو انعكاس لاختلاف المعرفة بين الشعبين، فعدد الأميركيين الذين يعرفون اسم رئيس الولايات أكثر من عدد الهولنديين، فإذا افترضنا أن هذا الاختبار هو وحيد البعد فمن المحتمل جداً أن يقيس هذا السؤال البناء العقلي الأساسي بشكل مناسب في كلا الدولتين. والنتيجة التي لا بد أن نصل إليها هي أن السؤال مفيد لأن النصين اللغويين للسؤال متماثلان تماماً ويقيس البنية العقلية الأساسية في كلتا الدولتين.

ومع ذلك فإن الاستنتاج بأن هذا السؤال صحيح ومفيد بسبب التماثل اللغوي بين النصين غير واضح، إذ إن هناك مشكلة عدم التماثل السيكولوجي بينهما: هل لهذا السؤال المعنى نفسه في كلا الدولتين. فالسؤال المماثل سيكولوجياً في هولندا قد يكون السؤال عن اسم الملكة أو رئيس الوزراء أو إحدى الشخصيات العامة الأخرى. إن تكييف النص الأصلي يفقده التماثل اللغوي إلا أنه يزيد من التماثل السيكولوجي. مع الأسف أن التكييف الذي يؤكد على التماثل السيكولوجي له مشكلاته أيضاً.

أول هذه المشكلات هي أن هذا التماثل يجب إقامته بالتجربة ولا يمكن الاكتفاء بافتراض وجوده، يضاف إلى ذلك أن طرح الأسئلة متباينة الصياغة وبلغات مختلفة سيقلل من إمكانية مقارنة النتائج بين الشعوب، مثال ذلك ما هي النتيجة التي نتوصل إليها إذا تبين أن 90٪ من الكهول في عينة الولايات المتحدة يعرفون اسم الرئيس، بينما 95٪ من الهولنديين يعرفون اسم الملكة. وإذا كانت العينة المدروسة واسعة فإن هذا الاختلاف قد يكون مهماً إحصائياً، إلا أن تفسير هذا الاختلاف قد يكون صعباً إن لم يكن مستحيلاً. فقد يعكس ذلك مثلاً اختلاف ظروف هؤلاء الأشخاص في وسائل الإعلام في بلديهما، بدل أن يكون سببه الاختلاف في بنية الاختبار أو في المعلومات العامة.

قد تكفي هذه المقدمة القصيرة للتأكيد على الموضوعات الرئيسة التي يتضمنها هذا الفصل، أولاً قد يتلاقى المنظور السيكولوجي واللغوي في أثناء الترجمة، وفي هذه الحالة تكون الترجمة دقيقة ومباشرة، لكنها قد تؤدي إلى صياغة مختلفة في اللغة المستهدفة. لكي نحصل على ترجمة جيدة للنص الأصلي يجب توافر خبرة جيدة باللغة الأصلية واللغة المستهدفة إلا أن ذلك غير كاف إذ لا بد من توافر منظور سيكولوجي حتى يمكن التوصل إلى أدوات (اختبارات) جيدة في اللغة المستهدفة (بيلينغ ولو، 2000، براكون وبارونا، 1991، برسلين، 1980، 1986،

كيسينجر، 1994؛ هامبلتون، 1994؛ جـ. هاركنس، 1998؛ ميريندا، 1994، فاليراند، 1989، فان دي فيجر، 2003).

ثانياً: هناك حاجة إلى إطار نظري نستطيع أن نحدد ضمنه بشكل دقيق ماذا نعني بتعبير "التماثل السيكولوجي".

نعرف في القسم الأول من هذا الفصل مصطلحين رئيسين يُستخدمان في تحديد التماثل وهما: الانحياز والتكافؤ. ونطبق في القسم الثاني هذين المصطلحين على الترجمة/ التكييف. ونعالج في القسم الثالث الطرق التي تعزز ملاءمة الترجمة/ التكييف. ونصف في القسم الأخير انعكاسات ذلك على استخدام الاختبارات.

يمكن اعتبار هذا الفصل كخلفية نظرية للإرشادات الخاصة في تكييف الأدوات التربوية والسيكولوجية وإحداث تكافؤ في الدرجات الموجودة في الجدول 3.1 في الفصل الأول. انطلاقاً من هذه الخلفية يصبح من السهل إدراك الأساس المنطقي للإرشادات وفهم المواقف التي تم تبنيها. وقد وردت الإشارة إلى الإرشادات كلما كان ذلك مفيداً.

الانحياز والتكافؤ

يختلف معنى كل من مصطلحي الانحياز والتكافؤ قليلاً في الأدبيات. تتضمن كلمة الانحياز وجود عوامل مزعجة (مؤذية)، يقال عن القياس إنه منحاز إذا اختلفت الدرجات المسجلة في أحد الاختبارات من لغة إلى أخرى بسبب وجود تباين غير مرغوب فيه. مثال: إذا أخذنا الترجمة السويدية - الإنجليزية في موضع الاختبار التالي: أين يحتمل أن تعيش الطيور ذات الوترات في أرجلها فإن الترجمة الراجعة من السويدية إلى الإنجليزية لعبارة (طير ذو وترات في أرجله) هي الطير ذو الأرجل السباحة هذا يعطي إشارة لمعرفة الإجابة تفوق ما يتضمنه النص الإنجليزي الأصلي (هامبلتون، 1994). مثال آخر في عمليات المسح الذي يتناول القيم في أوروبا، كانت

الدرجات للإخلاص والوفاء في إسبانيا مختلفة عن مجمل النتائج المسجلة في تلك الدولة. لدى التدقيق في الأمر تبين أن كلمة الوفاء التي استعملت في الترجمة إلى الإسبانية تضمن معنى الإخلاص الزوجي (هالمان، حزيران 1998).

يرتبط مصطلح التكافؤ ذهنياً بقياس أوجه الخلاف بين الأعراف، وما ينتج عن ذلك الانحياز. إن البند أو الأداة المنحازة ستعطي درجات غير متكافئة.

أصبح عدد التكافؤ تعبيراً عاماً يشير إلى عدم إمكان المقارنة بين الدرجات، وعملاً بهذا العرف فإننا نستعمل عدم التكافؤ كصفة مميزة لدرجات الاختبار التي تأثرت بفعل الانحياز الثقافي.

من وجهة النظر التي اتخذناها في هذا الفصل لا يعد الانحياز صفة جوهرية للاختبار وإنما هي نتيجة لتطبيق هذا الاختبار على مجموعة خاصة بغية الوصول إلى هدف محدد. وهي تشير إلى جميع أنواع العوامل المزعجة التي تعيق تفسير اختلاف الدرجات بين مجموعة وأخرى. يمكن فهم الانحياز بشكل أفضل من خلال القابلية للتعميم.

يمكن تعريف الانحياز بأنه التوافق غير المتوازن في مجال (ميدان) الملاحظة ودعائم التعميم. مثال على ذلك، لنفترض أننا طبقنا اختبار الامتداد الزمني (وهو الاختبار الذي يقيس اتساع الذاكرة القصيرة الأمد) على أطفال أميركيين وأطفال ريفيين من إفريقيا لم يحصلوا على تعليم مدرسة جيد ووجدنا أن الأطفال الأميركيين قد حصلوا على درجات أعلى. إذا فسّرت درجات الاختبار استناداً إلى عدد الأرقام التي يستطيع أطفال الفريقين الاحتفاظ بها في ذاكرتهم لمدة قصيرة، يمكن القول إن الاختبار غير منحاز؛ ذلك أن استخدام الأرقام في بعض المجالات مثل علم الحساب قد يعطي نتائج متماثلة الاختلاف. أما إذا اعتبرت علامات الامتداد الرقمي على أنها دليل على قدرات الذاكرة قصيرة الأمد (وهو التفسير الشائع لهذه الدرجات) فإن الاختبار يكون منحازاً على الأرجح. هناك إثبات في علم

النفس الشامل للثقافات، وهو أن اتساع الذاكرة قصيرة الأمد يختلف باختلاف الثقافات والأعراف (واغنر، 1981). ومع ذلك فإن اختلاف الحوافز بين الثقافات قد يكون له تأثير شامل على الدرجات في كثير من الاختبارات. إن الاختلافات التي شوهدت في اختبار الامتداد الرقمي قد لا تتكرر في الاختبارات التي تستخدم حوافز أكثر صحة من الناحية البيئية بالنسبة للأطفال الأفارقة الريفيين. لدى دراسة الانحياز سنحاول معرفة ما هي التبدلات التي تؤثر في أحداث الاختلافات المشاهدة بين الأعراق.

ثلاثة أنواع من الانحياز:

على الرغم من أن الانحياز قد ينشأ من مصادر عديدة فإن من الضروري التمييز بين عدة فئات. كثيراً ما تؤدي المصادر المختلفة إلى أنواع متماثلة من الانحياز. في اعتقادنا يوجد ثلاث فئات من الانحياز: انحياز البنية، انحياز المنهج وانحياز الموضوع (فان دي فيجر ولونغ، 1997 a، 1997 b، 2000؛ فان دي فيجر وبورتينغا، 1997؛ فان دي فيجر وتانزر 1997) (انظر الجدول 2.1).

الجدول 2.1

أنماط الانحياز

نمط الانحياز	الوصف
انحياز البنية	تداخل غير تام في البنية لدى الفئات الثقافية.
انحياز المنهج	مصطلح عام يشمل جميع عوامل الخلل الناتج عن أحد جوانب المنهج.
انحياز الأداة	أشكال الأداة التي تسبب اختلافات الدرجات في الإثنيات والتي لا علاقة لها بالبنية.
انحياز إداري	قصور التواصل بين منفذي الاختبار والخاضعين له.
انحياز الموضوع/ اختلاف عمل الموضوع	اضطراب الموضوع (مثل الترجمة السيئة).

انحياز البنية: يقصد بهذا الشكل من الانحياز التباين في البنيات الثقافية بين المجموعات المختلفة. يبين الجدول 2-2 نظرة إجمالية لأهم مصادر انحياز البنية. مثال ذلك أن العوامل التي تشكل البنية (مثل السلوك، المواقف، القواعد السلوكية) ليست متطابقة تماماً بين المجموعات المختلفة. إن أنصار نسبية المواقف التي تشاهد السيكولوجيات المحلية (اسينها، 1997) والسيكولوجيات الثقافية (كول، 1996، غرين فيلد، 1997 a، 1997 b، ميلر، 1997) يميلون للاعتقاد (فيما يتعلق بهذا الفصل) أن انحياز البنية هو القاعدة وليس الاستثناء في السيكولوجا عبر الثقافات.

الجدول 2-2

مصادر نموذجية لثلاثة أنماط من التحيز في تقويم عبر اللغات

نمط الانحياز	مصادر الانحياز
• انحياز البنية	• عدم التماثل في تعريف البنية عبر الثقافات. • تفاوت ملاءمة السلوك المرافق للبنية (مثال: مهارات ليس لها ذخيرة في ثقافة المجموعات).
• انحياز المنهج	• عدم إمكانية مقارنة العينات (مثال: يحدث هذا بسبب الاختلافات في التعليم، الدوافع). • تفاوت الألفة مع المواد المنبهة. • تفاوت الألفة مع إجراءات الاستجابة. • تفاوت طرق الاستجابة (مثال: الرغبات الاجتماعية، الدرجات القصوى، القبول). • تفاوت الظروف المحيطة بإجراء الاختبار، فيزيائية (مسجلات) أو اجتماعية (عدد الطلاب في الصف).
• انحياز البند	• ترجمة سيئة للبند/ بنود غامضة. • عوامل سيئة ذات علاقة بالبند.

ملاحظة: انظر (فان دي فيجر وتاتزر ١٩٩٧).

من الأمثلة على ذلك مفهوم الذكاء. تميل معظم اختبارات الذكاء إلى استخدام تعريف ضمني للذكاء يتألف من المحاكمة والتفكير المنطقي (كما هو الحال في اختبارات رافن) وبدرجة أقل من المعارف المكتسبة والذاكرة (كما هو الحال في مجموعات الذكاء مثل سلم فكسلر للذكاء عند الأطفال وسلم فكسلر للذكاء عند البالغين). نشاهد هذه العوامل أيضاً عندما يطلب من الأفراد بيان مميزات الشخص الذكي (ستينبرغ، كونوي، كيرتن وبيرستاين، 1981). إلا أن الدراسات في الأوساط غير الغربية بينت أن المفهوم الشائع للذكاء أوسع من ذلك ويتضمن بعض المظاهر الاجتماعية. مثال ذلك ما قالته أمهات كوكوه في كينيا من أن الطفل الذي يعرف مكانه في العائلة والسلوك هو الذي يتوقع منه مثلاً اتباع الطرق المناسبة في مخاطبة الآخرين، وأن الطفل الذكي هو الطفل المطيع الذي لا يسبب مشكلات (مندي - كاستل، 1974، سيجال، داسن، بيري وبورتينغا، 1987). وقد بينت دراسات مماثلة في زامبيا (سيربل، 1993) واليابان (أزوما وكاشيواجي، 1987) أن صفات الشخص الذكي تتجاوز ميدان المدرسة الذي يُعتمد عادة في الولايات المتحدة وأوروبا. ومثال آخر على الاختلاف في مضمون الذكاء يمكن أن يشاهد في أعمال (هو، 1996) حول طاعة الوالدين في الصين. وقد بين أنه بالمقارنة مع الغرب يميل الصينيون إلى تطبيق تعريف أوسع للذكاء. وأن الطاعة واحترام الأبوين هي عناصر تشاهد أيضاً في الدول الغربية، إلا أن المفهوم الصيني لاحترام الأبوين يشمل أيضاً العناية المادية بالأبوين عندما يتقدمان في السن ويحتاجان للمساعدة.

إن مشكلة اختيار العينات السيئ يضاعف عندما يتم العمل بوسائل مختصرة كما هي الحال في أكثر الأحيان. وقد شكا تريانديس (1978) قبل عدة سنوات من أن قياساتنا تتناول عينات قليلة من اهتمام البشر. ويقود هذا إلى ما يدعوه أمبرسون (1983) (نقص تمثيل البنية) إذ يتم اختيار عدد قليل من الموضوعات بسب تجانسها وتعد أنها تغطي بنيات واسعة. وعلى الرغم من أن هذه القضية ليست خاصة بميدان السيكولوجيا عبر الثقافات إلا أنها تبرز في هذا المجال بشكل

خاص. وعندما تُبدي مقارنة المجموعات الثقافية اختلافات واسعة، فإن القياسات الضيقة تبدي على الأرجح انحيازاً يعود إلى محدودية تمثيل البنية.

انحياز المنهج:

انحياز المنهج هو مصطلح عام لنوع ثان من الانحياز الذي يتضمن كل المتحولات المزعجة الناتجة عن عوامل خاصة بالمنهج وقد ابتكر هذا المصطلح لأن هذه العوامل تذكر عادة في القسم الخاص بالمنهج في الدراسات التجريبية (الإمبريقية).

هناك نوعان من انحياز المنهج. الأول هو (انحياز الأداة). يتضمن ذلك كل خصائص الأداة التي لا علاقة لها بهدف الدراسة، إلا أنها مع ذلك تسبب اختلافات في درجات الاختبار. إن السبب الأكثر شيوعاً لانحياز الأداة في الاختبارات الذهنية هو تآلف الأشخاص مع المنبهات والاستجابات (أو شكل الإجابة). يمكن العثور على إيضاح لذلك في دراسة قام بها سيربل (1979).

اهتم هذا الباحث بالمهارات الإدراكية/ الحسية عند الأطفال البريطانيين والزامبيين. وقد طلب منهم استنساخ صور أشخاص باستخدام الورق وقلم الرصاص، أوضاع اليد، والأسلاك المعدنية (التي هي شائعة في زامبيا). كما هو متوقع فقد تفوق الأطفال البريطانيون في الرسوم المعمولة بالورق وقلم الرصاص، بينما حصل الزامبيون على درجات أعلى بشكل واضح في الرسوم المعمولة بالأسلاك المعدنية، وكما هو متوقع أيضاً لم تسجل اختلافات باستعمال الوسائل الأخرى. وتفسير هذه المشاهدات باختصار على أنها ناجمة عن اختلاف التآلف مع الأجوبة.

يمكن للبيانات الخاصة بالشخصية والموقف أن تُبدي انحياز الأداة. مثال ذلك ما وجده (هوي وتريانديس 1989) من أن المكسيكيين كثيراً ما يحصلون على علامات مفرطة بالمقارنة مع الأميركيين من أصل أوروبي عند استعمال سلم مؤلف من 5 نقاط. ولا يشاهد هذا الانحياز عند استخدام سلم مؤلف من 5 نقاط تشير الفئة الثانية من انحياز المنهج إلى "الانحياز الإداري" ويقصد به اختلاف الدرجات

الناجم عن التعليمات وغير ذلك من مشكلات التواصل بين المختَبِرين والمختَبَرين. تحدث هذه المشكلات بشكل خاص عندما يستخدم المختبرون أو المختبرين لغة غير لغتهم الأم. قد يكون فقد المعلومات ناجماً عن عدم القدرة على التعبير عن الأفكار بلغة ثانية (غاس وفارونيس، 1991). كما أن عدم معرفة ثقافة المختبرين قد يؤي إلى انتهاك القواعد المحلية للمجاملة.

وقد عالج المختصون بعلم النفس والباحثون في المسوحات تأثير مميزات الشخص الذي يُجري المقابلة (كالجنس، العمر، العرق) على نتائج القياس. في مراجعة للدراسات السيكولوجية الخاصة بجنس المختبر على أداء الأطفال في اختبارات الذكاء استنتج "جنسن" (1980) أن الدراسات ذات المنهجية المناسبة قليلة. (مثال: لا توجد دراسات تعالج قضية عرق المختبر (المُختبر). مع ذلك إن البينات المتوافرة لا تشير إلى أن عرق المختبر ذو أهمية كبيرة. وقد درس الباحثون في المسوحات ما يُدعى بنظرية الإذعان (الاحترام). فقد وجد كوتر، كوهين وكولتر (1982) أن الأشخاص أكثر ميلاً لإبداء مواقف إيجابية تجاه جماعة ذات ثقافة معينة عندما تجري مقابلتهم من قبل أحد أفراد تلك الجماعة (رييس، دانيلس، شوميكر، شانك وهو، 1986). مع ذلك فإن أهمية التأثيرات العائدة لصفات الباحث الذي يجري المقابلة تبدو قليلة ومتضاربة في مختلف الدراسات (سنجر وبريس، 1989).

قد يكون لانحياز الأداة تأثير شامل على درجات الاختبار، عن تأثير الاختلافات في التعليم أو تعرض المختبر لاختبارات سابقة قد يؤثر على الدرجات في بعض الموضوعات ذات العلاقة بالمدرسة. من المرجح أن هذه الاختلافات ستؤثر على معظم الموضوعات أو جميعها، وهكذا فإن انحياز المنهج قد تكون نتيجة لمقارنة قياسات نفسية، إذا عممت هذه الدرجات على الأفراد خارج نطاق المدرسة. إذاً قد ينتج عن انحياز المنهج اختلاف في الدرجات داخل المجموعات نفسها لا علاقة له بالبنية وإنما يعود إلى خداع القياسات. إن المعنى المتضمن لهذا الأمر خطير

منتدى سور الأزبكية WWW.BOOKS4ALL.NET https://twitter.com/SourAlAzbakya

فالباحث أو الممارس الذي يقارن الدرجات عبر الثقافات (سواء تم ذلك بشكل صريح عن طريق اختبار هذه الاختلافات إحصائياً، أو بشكل ضمني عن طريق تطبيق جدول معياري للفرد في الثقافات المختلفة) سوف يحتاج للاختيار بين تفسيرين متنافسين هما: الاختلافات المشروعة عبر الثقافات وانحياز المنهج. وغالباً ما يصعب الاختيار بين هذين التفسيرين بسبب نقص البينات التي تؤكد أو تنفي أياً منهما، وهكذا فإنه من المهم بالنسبة لمصممي الاختبارات أن يعترفوا بأهمية انحياز المنهج ويحاولوا التقليل من تأثيره إلى أبعد حد ممكن (هاميلتون، 1994؛ انظر أيضاً سيرغي، باتسولا، هامبلتون، الفصل الرابع في هذا الكتاب، وتهدف جميع التوجيهات الإدارية إلى تحقيق ذلك).

انحياز البند (اختلاف عمل البند)

يشير هذا النوع من الانحياز إلى الأخطار التي تؤثر على صحة البنود فقط، بينما يتناول انحياز البنية والمنهج المظاهر العامة للاختبار. في البدء استخدم تعبير انحياز البند (كليري وهيلتون، 1968). بعد أكثر من ثلاثة عقود من التطور المهم في القياسات النفسية لاكتشاف الموضوعات غير السوية (اكريكان، 2002؛ هولاند ووينر، 1993: ميلساب وافرسون، 1993؛ سيرغي وآلالوف، 2003)، تم استبدال هذا المصطلح "اختلاف عمل البند"؛ كان هناك شعور بأن انحياز البند يتضمن معنى الابتعاد عن المعيار الأوروبي/ الأميركي الذي كان وما يزال أكثر مجموعة معيارية مستخدمة في الأبحاث في الولايات المتحدة. إلا أننا نتمسك بالمصطلح الأصلي لأنه يؤكد على العلاقة الوثيقة مع الأنماط الأخرى من الانحياز ويشير إلى مظاهرها الأساسية: إنه تهديد لصحة البند ويمنع المقارنة المباشرة للدرجات.

إن أهم أسباب انحياز البند هو الترجمة السيئة واختلاف المعاني الضمنية للكلمات. مثال ذلك أنه استناداً إلى قاموس "وبستر" الأميركي فإن العدوانية "تتجلى على شكل تصميم واضح واستعداد للخصام" بينما قاموس اوكسفورد البريطاني

يعطي المعنى الأول للعدوانية بأنه "عمل أو ممارسة الهجوم دون تربص وبشكل خاص البدء بالحرب أو النزاع". يأتي المعنى السابق في القاموس الأميركي في المرتبة الثالثة. من المهم الإشارة إلى أن العدوانية في اللغات الأخرى كالألمانية والفرنسية والهولندية وغيرهم من اللغات هي أقرب إلى معناها من التعريف البريطاني وليس الأمريكي.

هناك مجموعة مذهلة من التعاريف والتقنيات الإحصائية التي تم اقتراحها لتعريف انحياز البند، إلا أن هناك في الوقت الحاضر ميلاً للتقارب في هذا المجال أكثر مما كان من قبل عشر أو عشرين عاماً. وقد أصبح أحد التعاريف ومجموعة صغيرة من التقنيات الإحصائية مقبولة على أنها الأكثر ملاءمة. ومن المهم في التعاريف المتداولة أن انحياز البند مشروط بمستوى الاستعداد أو السمة. ماذا نعني بذلك؟ لنعد إلى المثال المتعلق برئيس الولايات المتحدة، إن الدرجات العالية التي سجلها الطلاب الأميركيون في هذا البند تعكس اختلافاً حقيقياً بين المواطنين الأميركيين والهولنديين، ولا نريد أن نتجاهل هذا الاختلاف على أنه انحياز أو على أنه "تأثير معاكس" إذا استخدمنا مصطلحاً آخر يستخدم للتعبير عن الاختلافات غير الصحيحة بين الثقافات.

لتحليل مصطلح انحياز البند نقوم بتقسيم كل من العينتين إلى عدد من المجموعات بحسب الدرجات (تدعى هذه العملية التكييف الشرطي) Conditioning في المجموعة الأولى للأميركيين نضع جميع الأشخاص الذين حصلوا على درجة (1) في الاختبار، وكذلك الأمر بالنسبة لمجموعة الهولنديين التي تضم جميع الأشخاص الذين حصلوا على درجة (1). وتضم المجموعة التالية جميع الأشخاص الذين حصلوا على درجة (2)، ونقوم بالشيء نفسه بالنسبة لبقية الدرجات، يسمح لنا هذا الإجراء القيام بتحليل مفصل؛ فبدلاً من مقارنة متوسط الدرجات يصبح بإمكاننا مقارنة الأميركيين والهولنديين الذين حصلوا على المستوى نفسه من الدرجات (أي أننا نجري المقارنة المشروطة بمستوى الدرجات). يبين الاختبار الخاص برئيس الولايات المتحدة وجود انحياز البند إذا كان الأميركيون والهولنديون الذين لديهم درجة متماثلة في مجموع الاختبارات لم يحصلوا على المستوى نفسه من الدرجات في هذا

البند. وبصورة أدق نقول بوجود انحياز البند إذا كان الأشخاص الذين لديهم المستوى نفسه في اختبار البنية (وهم الأشخاص الذين لديهم درجة مماثلة في مجمل الاختبارات) لم يحصلوا على الدرجة المتوقعة نفسها في هذا البند (ويقصد بذلك متوسط الدرجة في هذا البند) (هولاند وثاير، 1988، 1997 b، 2000).

لقد وصف "ميلينبرغ" (1982، انظر أيضاً غلوستر ومازور، 1998) الفرق بين الانحياز المتساوق وغير المتساوق. يقال: إن البند منحاز بشكل متساوق إذا كان الاختلاف في مستوى أداء ثابتاً في جميع المستويات تقريباً (يقصد بذلك أنه في كل مجموعة من الدرجات يكون أداء الأميركيين المختبرين أفضل من أداء الهولنديين بالمقدار نفسه تقريباً). ويقال بوجود انحياز غير متسق إذا كان حجم الفارق يختلف بشكل مضطرد بين المجموعات (مجموعات الدرجات). مثال ذلك في المجموعات التي حصلت على درجات منخفضة كان عدد الأميركيين الذين يعرفون اسم رئيسهم أقل من عدد الهولنديين الذين يعرفون اسم الملكة، إلا أن الفارق يتضاءل تدريجياً عند المجموعات ذات الدرجات الأولى.

مستويات التكافؤ:

يتحدى الانحياز مقارنة الدرجات التي يتم الحصول عليها في المجموعات المختلفة. من الناحية التقنية يحدد الانحياز تكافؤ الدرجات. ولكي نحدد نتائج الانحياز عند مقارنة الدرجات نبين هنا أربعة أنماط من التكافؤ مرتبة بحسب صلاحيتها للمقارنة حسب الإثيات (فان دي فيجر ولونغ 1997 a، 1997 b، 2000؛ فان دي فيج وبورتينغا، 1997، انظر الجدول 2-3). أول هذه الأنماط يدعى عدم تكافؤ البنية. يتميز هذا النمط بانعدام القابلية التامة للمقارنة، كما لو قارنا البرتقال مع التفاح". هذا النمط من التكافؤ هو نتيجة انحياز البنية. يستحيل إجراء مقارنة الدرجات عبر الثقافات استناداً إلى عمليات/ فعاليات غير تامة أو غير ملائمة.

الجدول 2-3

أنماط التكافؤ

نوع التكافؤ	الوصف
تكافؤ البنية	تقيس الأداة بنيات مختلفة في ثقافتين مختلفتين (مقارنة التفاح مع البرتقال).
تكافؤ البناء/ التكافؤ الوظيفي	تقيس الأداة ذات البناء السيكولوجي عبر مجموعات ثقافية مختلفة.
تكافؤ وحدة القياس	تشتمل الأداة على ذات وحدة القياس وأصل مختلف عبر المجموعات الثقافية المختلفة.
تكافؤ المقياس/ تكافؤ الدرجات التام	تشتمل الأداة على ذات وحدة القياس وذات الأصل عبر المجموعات الثقافية المختلفة.

يُعرف النمط الثاني من التكافؤ بأسماء مختلفة، أكثرها شيوعاً هو التكافؤ البنيوي والتكافؤ الوظيفي. يترافق هذا النمط بصنف من الإجراءات التي تُستخدم لإقامة تطابق في البنيات بين المجموعات كما هو معمول به في أحد اختبارات القياس النوعية. بشكل عام يتطلب هذا الشكل من التكافؤ أن تكون نماذج العلاقات بين المتغيرات هي نفسها في كل واحدة من المجموعات. يُستخدم هذا النمط من التكافؤ في العديد من مشروعات تكييف الاختبارات (الارشادات 2).

عندما يجري تهيئة ترجمة اختبار ما إلى لغة أخرى تظهر مسألة صدق البنية: هل يقيس النص الأصلي والنص المترجم للاختبار البنية النفسية ذاتها؟ يكون هذا التساؤل مهماً عندما يُترجم الاختبار حرفياً. في مثل هذه الحالات غالباً ما يستخدم تحليل العامل المؤكد أو تحليل العامل الاستكشافي متبوعاً بتبادل الهدف لدراسة تماثل عوامل البند عبر ثقافات سكانية (لمزيد من التفاصيل انظر: ايرن،

شافلسون وماثن، 1989؛ ليتل، 1997؛ فان دي فيج ولونغ، 1997 a، 1997 b، 2000؛ واتكينز، 1989). إن تماثل محتويات العامل في كل بند ينظر إليه كشرط أساسي للتكافؤ البنيوي، وينطبق هذا الأمر على الاختبارات المترجمة حرفياً. إذا حدثت تغييرات في اللغة المترجم إليها خلال عملية الترجمة من أجل الحصول على بنية مناسبة (أي عندما يتم إجراء أي تعديل) يجب أن تظهر العوامل السابقة نفسها، إلا أنه لا يمكن توقع حصول تطابق في كل واحد من البنود مع الآخر، على سبيل المثال هناك أكثر من أربعين ترجمة لكل من سبيلبرغر، غوسج، ولوشينيس "الاختبارات الشخصية لسمات القلق" (STAI) لم تكن الغاية المهمة من أكثر هذه الترجمات إنتاج ترجمة حرفية للنص الإنجليزي وإنما إعداد اختبار أداة قادرة على تقييم القلق بشكل مناسب في الثقافة التي يجري الترجمة إليها. إذا أخذنا هذا الهدف بعين الاعتبار يصبح العالم الخاص بتقنيات التحليل أقل ملاءمة ويصبح اختبار شبكة النواميس المهيمنة على حقل البند أكثر أهمية (كرونباخ وميهل، 1955). باختبار هذا الإجراء النموذج المتوقع للروابط العالية مع المقاييس الأخرى للقلق المتوافرة سلفاً في اللغة الهدف، كما يقوم باختبار الروابط القليلة أو المعدومة مع المقاييس الخاصة بالبنية غير ذات الصلة بالموضوع.

أما المستوى التالي من التكافؤ فيدعى "قياس وحدات التكافؤ"، يتطلب هذا المستوى أن يكون المقياس في كل مجموعة له ذات المسافة (أي أن القياسات هي نفسها على مختلف المستويات). إن مقياسين يبديان وحدات قياسية متكافئة عندما يكون لها ذات الوحدات القياسية وكلها ذات مصدر مختلف. وهذا هو أدنى مستوى من التكافؤ وفيه يمكن مقارنة مستوى الدرجات بشكل صحيح ولو كان ذلك مع بعض القيود، فالفوارق الإفرادية التي تُشاهد في المجموعة A يمكن مقارنتها بالفوارق الإفرادية في المجموعة B، مثال ذلك إذا حقق اختبار انبساط الشخصية هذا النمط من التكافؤ أمكن الإجابة عن التساؤل عما إذا كانت اختلافات الجنس (ذكر أو أنثى) في ذلك الاختبار متماثلة عبر المجموعات الثقافية المختلفة، ومع ذلك نظراً لاحتمال

وجود اختلاف في مصدر المقياس لا يمكن إجراء مقارنات بين الدرجات عبر المجموعات. وهكذا يصبح من المستحيل القول بأن إحدى المجموعات تحصل على مستوى أعلى من غيرها في هذا الاختبار (اختبار انبساط الشخصية) أو أن الشخص X من المجموعة A هو أكثر انبساطاً من الشخص Y من المجموعة B عندما تكون وحدات المقياس متكافئة.

إن الاختلافات الأساسية في مستوى الدرجات يمكن دراسته فقط عندما تبدي الدرجات تكافؤ درجات أعلى "تكافؤ المقياس/ تكافؤ الدرجات التام".

إن القياسات التي تُبدي هذا النمط من التكافؤ تملك ذات وحدات القياس وذات المصدر في كل المجموعات. في هذا النمط من التكافؤ تستطيع الدرجات تجاوز الحدود الثقافية دون أي مشكلات كما يمكن مقارنتها بشكل سليم لدى أشخاص من ثقافات وأعراق مختلفة.

التكافؤ والانحياز:

إن التكافؤ والانحياز مرتبطان بشكل وثيق، حتى إن بعض المؤلفين يستخدمون هذين المُصطلحين للدلالة على المعنى ذاته، بعد التعرف على هذه المصطلحات أصبح من الممكن دراسة العلاقات بينها بشكل مفصل. قد يُخفض الانحياز أو لا يُخفض مستوى التكافؤ في المقارنات كما هو مبين في الجدول 2-4 وسيُحدد مستوى التكافؤ نمط المقارنة الذي يمكن اتخاذه (الإرشادات 21).

الجدول 2-4

تأثير الانحياز على مستوى التكافؤ

	مستوى التكافؤ		
- نمط الانحياز	البنية	بند القياس (أ)	المقياس (ب)
- انحياز البنية	نعم	نعم	نعم
- انحياز المنهج المتسق	لا	لا	نعم
- انحياز المنهج غير المتسق	لا	نعم	نعم
- انحياز البند المتسق	لا	لا	نعم
- انحياز البند غير المتسق	لا	نعم	نعم

إن مقارنة الدرجات المباشر (مثل أفراد المجموعة A أكثر انبساطاً في المتوسط من أفراد المجموعة B) يتطلب مستوى أعلى من تكافؤ البنية (مثال: هل يعد الانبساط على أنه بنية متماثلة في أفراد المجموعتين A وB) إن انحياز البنية هو أهم التحديات التي تواجه مقارنة الدرجات لأنه يستخدم شكلاً من عدم التكافؤ الذي يحول دون أية مقارنة عبر الثقافات؛ لهذا السبب قد يبدو عدم التكافؤ غير مرغوب به، ومع ذلك فإن هذا الوضع قد يكون نقطة بدء لاستشكاف الفروقات المهمة عبر الثقافات التي تتجاوز انحياز البند أو المنهج وقد تشمل المفهوم التكويني للبنية (غرينفيلد، 1997 a؛ بورتينغا وفان دير فلير، 1988).

إذا كنا مهتمين بمعرفة ما إذا كان الاختبار يقيم بعض السمات أو القدرات في المجموعات المختلفة فيكفي إقامة تكافؤ بنيوي.. في هذه الحالة قل أن يكون انحياز المنهج ذا أهمية أساسية. تأخذ التحليلات الإحصائية للتكافؤ البنيوي بعين الاعتبار العلاقات المتبادلة (الروابط) في المقام الأول ولا يؤثر انحياز المنهج على هذه الروابط؛ لذلك فإن هذه الإحصائيات لا تتأثر بانحياز البنية.

عندما يتم تحليل درجات البند تصبح الصورة أكثر تعقيداً. وقد اقترح لمعالجة موضوع التكافؤ البنيوي تحليل عاملي استكشافي يتلوه تدوير الهدف. إن الاختلافات في محتوى العامل عبر المجموعات تؤخذ على أنها دليل على انحياز البند (إذا كانت قلة من البنود منحازة) أو غياب التكافؤ البنيوي (إذا كانت عدة بنود منحازة). إذا أبدى أحد البنود انحيازاً متمثلاً فمن المحتمل أن تبقى الروابط بين البند مع غيره من البنود ثابتة. لن يُشار إلى هذا البند على أنه مشبوه في تحليل عاملي استكشافي. مع ذلك فإن التحليل العاملي الاستكشافي سيكون دقيقاً في الانحياز غير المنظم؛ لأن هذا النمط من الانحياز يُرجح أن يؤدي إلى نماذج متباينة من الروابط عبر المجموعات الثقافية بين هذا البند وغيره من البنود. تدل الكتابات حول انحياز البند أن الانحياز المتسق أكثر شيوعاً من الانحياز غير المتسق. وهكذا فقد يكون تحليل عامل الاستكشاف متبوعاً بتدوير الهدف إجراءاً كافياً لتحديد انحياز البند عندما يكون القصد تكافؤ البنية فقط. عند مناقشة الأنماط القياسية للتكافؤ (وحدات قياس وتكافؤ تام في الدرجات) يجب أن يؤخذ بعين الاعتبار التمييز بين الانحياز المتسق والانحياز غير المتسق. لا يتعارض الانحياز المتسق مع التكافؤ المقاس بالوحدات؛ لأنه مع هذا المستوى من التكافؤ لا يمكن مقارنة الدرجات مباشرة عبر الثقافات: إن إضافة رقم ثابت إلى جميع الدرجات ضمن مجموعة واحدة لا يسيء إلى هذا النمط من التكافؤ. ومع ذلك فإن الانحياز المتسق سوف يؤثر على فاعلية المقارنة بين الدرجات التي تبدي تكافؤ مقياس ما من جهة ثانية سوف يؤدي الانحياز غير المتسق إلى عدم تكافؤ في القياس وفي الدرجة النهائية؛ لأن هذا النمط من الانحياز يقضي على تماثل وحدة القياس عبر المجموعات.

انحياز العينة:

إن جميع أشكال الانحياز التي استعرضناها في هذا الفصل تتناول التفاوت في تقييم درجات الاختبار على البشر وعلى مجال معين من السلوك في الثقافات

المختلفة. إلا أن نمطاً واحداً من الانحياز لم يتم ذكره وهو انحياز العينة أو عدم قابلية العينات للمقارنة. هذا يتعلق بتباين العينات في قدرتها على تمثيل ثقافة السكان الذي تؤخذ منهم العينات. مثال ذلك أنه في كثير من الدراسات تم أخذ العينات من طلاب الجامعات إلا أن دخول الجامعة في بعض البلدان يعتمد في الدرجة الأولى على الأداء المدرسي فيما يكون الوضع الاقتصادي والاجتماعي للأهل في بلدان أخرى هو المقياس الأول في قبول الطلاب في الجامعة. وهكذا إن هنا اختلافات في نظام العينات فيما يتعلق بمواصفات الخلفيات مما يعلل أي تباين يلاحظ في الدرجات بالإضافة إلى الصفات الثقافية للمجموعة المعنية. إن أحد الشواهد الموثقة في الكتابات عبر الثقافات تتعلق بالنتائج المعرفية للأمية. ففي كثير من الدراسات القديمة في هذا الخصوص انقلبت المقارنة بين الأميين وغير الأميين إلى مقارنة بين الأشخاص المنتسبين إلى المدرسة وغير المنتسبين لأن الانتساب إلى المدرسة والتعلم (معرفة القراءة والكتابة) مفهومان مختلفان (يتعلم المرء القراءة والكتابة في المدرسة). وهكذا فإن الاختلاف في الانتساب للمدرسة يوفر تعليلاً بديلاً لمعرفة القراءة والكتابة عند شرح الاختلافات. وقد درس سكريبنر وكول (1981) معرفة القراءة والكتابة في ليبيريا كما درسها بينت في كندا (1991) فوجدوا أن القراءة والكتابة تكتسب عن طريق التعليم غير الرسمي (خارج المدرسة) ومن الجدير بالاهتمام أن كلاً من تلك الدراسات وجدت أن النتائج المعرفية لعدم التعليم المدرسي كانت ضئيلة.

خيارات الترجمة/ تكييف الاختبارات:

جرت المناقشة من قبل أن المعايير اللغوية والنفسية للترجمة الجيدة لا تلتقي دوماً. فالكلمة والجملة وأي نص في أي اختبار تقييم قابل للترجمة بشكل جيد إذا كان تحويل النص من اللغة الأصلية إلى اللغة الهدف يحتفظ بجميع ميزات النص الأصلي. بعبارة أخرى يُعد النص قابلاً للترجمة إذا اتفقت الاعتبارات اللغوية والنفسية على تعريف الترجمة المثلى. تُركز الاعتبارات اللغوية على تساوي دلالات

الألفاظ (هل تعطي الترجمة المعاكسة النص الأصلي؟) وقابلية الفهم، القراءة والأسلوب. أما الاعتبارات النفسية فتشير إلى غياب الأنماط المختلفة من الانحياز التي ذكرت في الفقرات السابقة وبالتالي فهي تشمل اللغة الواقعية فيما تركز الاعتبارات اللغوية على مظاهر النص، تنظر الاعتبارات النفسية إلى الاختبار في نطاقه الثقافي الواسع. تتوافر ثلاثة خيارات للترجمة استناداً إلى قابلية الاختبار للترجمة. (هناك عدة فصول في هاركنس، فان دي فيجر وموهلر، 2003 التي تعالج موضوع الترجمة وتصميم الاستبيانات متعددة اللغات).

خيار المطابقة:

تكون عملية الترجمة سهلة إذا اتضح أن الترجمة اللغوية المناسبة كانت ملائمة أيضاً في الناحية النفسية. ترجمة مثل ذلك ستكون ترجمة حرفية في الغالب ولا تتطلب تغييرات كبيرة في الكلمات. وقد أطلق فان دي فيجر ولونغ (1997 b) على هذا الخيار اسم المطابقة لأن النص في اللغة الأصلية يمكن تطبيقه ببساطة في سياق ثقافي آخر. من المحتمل إلى حد كبير أن دراسة طرق الترجمة التي يُشار إليها في الصحف عبر الثقافات، مثل مجلة "علم النفس عبر الثقافات" ستبين أن معظم المقارنات المنشورة تعتمد على الترجمة الحرفية. إن هذا الخيار هو الأكثر استعمالاً في البحوث التجريبية لسببين اثنين. الأول هو أن هذه الترجمة سهلة التطبيق مما يجعل تكاليفه معقولة، إضافة إلى ذلك فإنها تحتفظ بكل إمكانيات المقارنة المتعلقة بتكافؤ القياسات، إلا أن خيار المطابقة له عيب مهم: فهو صالح للاستعمال فقط عندما يكون الانحياز (لا سيما انحياز البنية وانحياز المنهج) أمراً مستبعداً. من المؤسف أن كثيراً من الدراسات عبر الثقافات تستخدم هذه الطريقة (السريعة والسيئة) في الترجمة الحرفية دون الأخذ بعين الاعتبار أخطار المقارنة عبر الثقافات.

خيار التكييف:

الخيار الثاني يدعى التكييف. يُقصد بالتكييف في مصطلحاتنا الترجمة الحرفية لبعض المحرضات (المنبهات) وتبديل بعضها الآخر بحيث تضاعف من ملاءمتها لثقافة اللغة المستهدفة. وقد أصبح التكييف مصطلحاً عاماً في هذا الكتاب وفي غيره من المنشورات المتعلقة بعلم النفس يشار به إلى ترجمة الاختبارات. وقد كان اختيار هذا المصطلح مقصوداً، لأنه يؤكد على قصور الترجمة الحرفية وعلى الحاجة على الأقل إلى النظر في ملاءمة هذه الترجمات من الناحية النفسية إن لم يكن تغيير المظاهر البارزة للاختبار. تظهر الحاجة للتكييف في الاختبارات التي تبدي قابلية متوسطة للترجمة: قد يكون بالإمكان ترجمة بعض معالم الاختبار كالتعليم والأمثلة والتمارين مباشرة إلى اللغة المستهدفة إلا أن الكلمات قد تكون بحاجة للتبديل، قد تكون هذه التعديلات مطلوبة للتعامل مع مختلف أشكال الانحياز كما بحثنا ذلك مسبقاً. يمكن أن نجد أمثلة على ذلك في كتابات STAI لوغلانسمان، سكافنر وسبيلبرغر 1981؛ سبيلبرغر وغيره، 1970). "تم تعديل قائمة مينيسوتا لدراسة الشخصية" لتلائم العديد من الثقافات فقد عدّلها على سبيل المثال لوشيو، ريس - لاغونس وسكوت (1994) لتلائم المكسيك وعدّ لها تشونك (تشونك ولونغ، 1998) لتلائم الصين. ومن الأمثلة الأخرى أن ليو عدّل (اختبار القدرات المعرفية) وهو اختبار تم تطويره في الولايات المتحدة كي يُصبح صالحاً للاستخدام في تشخيص الخرف عند السكان الصينيين ذوي المستوى المنخفض من التعليم الرسمي.

لا يهدف واضعو الاختبارات المكيف إلى إجراء مقارنات عبر الثقافات، وإنما يهدفون على الأرجح إلى تغطية بنية محددة في إحدى المجموعات الثقافية. وهكذا يُعد الهدف المحدود للمقارنة عبر الثقافات أمراً مفروغاً منه (الإرشادات 12).

خيار التجميع:

الخيار الثالث هو خيار التجميع. يستخدم هذا الخيار عندما تكون الأداة (الاختبار) عسيرة الترجمة. عندما تكون الترجمة الحرفية للأداء (الاختبار) غير

ممكنة لأسباب متعلقة بانحياز البنية أو المنهج وكان تكييف الاختبار لا يغطي البنية بشكل ملائم، فقد يعمد واضع الاختبار إلى تطوير/ جمع اختبار جديد تماماً في اللغة المستهدفة.

قام سيربل (1993) بدراسة مفهوم الذكاء عند الأفراد في زامبيا ووضع اختباراً مستنداً إلى هذا المفهوم يُعد مثالاً لاختبار التجميع. والمثال الآخر على هذا الخيار هو دراسة تشرش (1987) للشخصية في الفليبين التي قادته إلى وضع توجيهات لوضع اختبار للشخصية أكثر ملاءمة من الناحية الثقافية. يمكن الإشارة في هذا المجال إلى تشونك ورفاقه (1996) الذين وضعوا قائمة لتقييم الشخصية "الصينية" وهو اختبار يحتوي على العديد من أبعاد الشخصية المحلية مثل "الوجه" و"الانسجام". في كل هذه الأمثلة حاول الباحثون وضع صورة ملائمة للتكوين النفسي مستمدة من المفاهيم المحلية التي درسوها قبل وضع الاختبار. من وجهة نظر التكافؤ لا يؤمن خيار التجميع أي مجال لمقارنة الدرجات بشكل مباشر. مع ذلك فإن هذه الدراسات وثيقة الصلة بعلم النفس عبر الثقافات إذ تبين أن المفهوم الغربي أو الاختبار لا يمكن تطبيقه على بنية بعض الثقافات الأخرى. إن هذه الدراسات فعالة في تحقيق الغاية الأساسية من دراسة سيكولوجيا الثقافات وعبر الثقافات: وهي تحديد الانحياز الغربي في النظريات والاختبارات الحالية السائدة في ميدان علم النفس.

إن انتقاء أحد خيارات الترجمة أي المطابقة أو التكييف أو التجميع أمر مهم في أي مشروع بسبب مضامينه. فإذا كانت الغاية هي تحقيق تكافؤ قياسي (وحدة قياس أو درجة تامة) فإن خيار المطابقة هو الخيار الأفضل. على الرغم من أن بعض النماذج الإحصائية تسمح بمثل هذه المقارنات باستخدام اختبارات مكيفة. يفترض خيار المطابقة عدم وجود انحياز في البنية أو المنهج (انحياز البند يمكن معالجته بعد ذلك بحذف البند). في الوقت الذي يتعرض فيه العلماء إلى ضغوط كبيرة من

أجل نشر أبحاثهم فإن خيار المطابقة قد يصبح بسهولة الخيار الأسهل تناولاً لأنه يجمع بين الجهد القليل لإنجاز الترجمة مع التعهد الضمني بالتوصل إلى تكافؤ معياري. وهذا أمر مؤسف من وجهة نظر عبر الثقافات؛ لأن مثل هذه الممارسة قد تؤدي بسهولة إلى تطبيق اختبارات غير مناسبة ثقافياً، وإلى تقييم خاطئ للتباينات السيكولوجية بين أبناء الثقافات المختلفة (بورتينغا، 1975).

إننا نتفق مع المبدأ الأساسي في مختلف الإرشادات (مثل 20 و11 -(2.4) هامبلتون، 1994، فان دي فيجر وهامبلتون، 1996؛ انظر أيضاً هامبلتون، الفصل الأول في هذا الكتاب. إن مهمة الباحثين هي تبيان ملاءمة اختباراتهم. إن هذا الاقتراح ينحرف عن الممارسة الشائعة وفيها يقع عبء الإثبات على أكتاف الذين يستخدمون الاختبار.

عوضاً عن تضخيم ملاءمة الاختبار للمقارنة عبر الثقافات يمكن أيضاً زيارة صدقه البيئي. يمكن تحقيق ذلك بوساطة خيار التجميع وبدرجة أقل بواسطة خيار التكييف. والخلاصة يبدو أن تضخيم ملاءمة الاختبار سواء للمقارنة عبر الثقافات أو صدقها البيئي في إطار ثقافي محدد قد يؤدي إلى مقاربة مختلفة في عملية الترجمة/ التكييف.

تعزيز صدق الاختبار:

يتطلب التعامل بشكل فعال مع انحيازات اللجوء إلى مقاربة تحريضية. إن الدفاع عن هذه المقاربة كما عُبّر عنها في الفقرات السابقة يبقى ناقصاً دون إلقاء نظرة إجمالية على المناهج والإجراءات لتعيين وضبط/ مراقبة وحتى حذف الانحيازات. نحاول في هذا القسم إعطاء مثل على هذه النظرة الإجمالية ومع ذلك يتطلب الأمر الإيضاح أنه من غير الممكن تقديم قائمة شاملة، إذ إن كل مشروع لتكييف اختبار جديد سوف يصادف بعض الانحيازات الخاصة بذلك المشروع. من جهة أخرى سيكون هناك العديد من النقاط المتكررة. إن هذه المناقشة للمناهج والإجراءات ترتكز على الموضوعات المهمة التي نراها مهمة.

عندما يكون هناك شك بأن انحياز البنية والسلوك المرافق للبنية غير متماثل عبر المجموعات الإثنية يصبح من المستحيل ابتكار أي أداة (اختبار) تعطي درجات قابلة للمقارنة بين السكان من ثقافات مختلفة. من المهم تعيين المدى الذي وصل إليه التشابك (الإرشادات 2). إن القول بأن الولاء للأبوين أعلى (أو أقل) في الصين مثلاً مما هو في المملكة المتحدة قول مضلل دون الإشارة إلى أن البيئة الصينية (بنية الاختبار الصيني) غير كاملة إذا كانت تعتمد على أدوات غربية تهمل المظاهر المادية للولاء للأبوين، إذا كانت النتائج المعتمدة على التعريف الصيني يصبح الاختبار مفرط الشمولية في المملكة المتحدة.

هنالك على الأقل طريقتان للتعامل مع المشكلة. الأولى هي إعادة تحديد البنية التي يُراد قياسها بطريقة تتضمن الإشارة إلى فرط الشمولية أو نقص الشمولية، فعوضاً عن المقارنة بين الولاء الأبوي بشكل عام يمكن أن نصف نتائجنا على شكل المظاهر المادية وغير المادية للولاء الأبوي. والطريقة الثانية هي تطبيق تقنيات إحصائية خاصة لمعالجة مجموعات متباينة من الحوافز مثل نظرية الإجابة على البند (فيشر ومولينار، 1995؛ هامبلتون وسواميناثن، 1985، هاملتون، سواميناثن وروجرز، 1991؛ فان دير ليندن وهامبلتون، 1997) أو صياغة التبادل البنيوي (بولن، 1989؛ بيرون. 1998؛ ماركوليدس وشوماخر، 1969 (الارشادات 11).

إن المبدأ الأساسي هو تفكيك عالم كبير إلى واحد أو أكثر من الحقول الثانوية المشتركة عبر الثقافات مع الاحتفاظ في الوقت نفسه في كل مجموعة ثقافية على العلاقات بين هذه الحقول الثقافية الفردية.

إن إحدى الطرق لتجنب انحياز البنية هي اللامركزية (ورنر وكامبل، 1970). يُستخدم هذا المفهوم في الوقت الحاضر بأساليب مختلفة. إن المعنى الأصلي لتعبير اللامركزية يُقصد به تطوير اختبار بعدة لغات في آن واحد. وخلافاً لما هو قائم في معظم مشاريع الترجمة/ التكييف التي يتم فيها ترجمة أداة موجودة سابقاً إلى لغة

أخرى، تؤلف مجموعة تضم واضعي الاختبارات بالإضافة إلى ممثلين للغات المستهدفة. في هذه الحالة تصبح تعريفات البنية ومظاهر البنية مثل التعليمات والبنود المنحازة نحو مجموعة ثقافية معينة قابلة للتمييز من قبل المطورين من الثقافات الأخرى.

يمكن حذف الكلمات أو الجمل التي تشير إلى عادات أو معارف ثقافية معينة وتسيء إلى قابلية ترجمة الأداة والاستعاضة عنها بما يقابلها في التعابير القابلة للاستخدام عالمياً. وبالمقابل قد يقرر الباحثون أن ترجمة متكافئة غير ممكنة وأن من المتعذر إنشاء اختبار يسمح بمقارنة كمية. إن الأمثلة على مثل هذه المقارنة نادرة على الأرجح بسبب صعوبتها؛ ولأن الباحثين الذين استخدموا أحد الاختبارات من قبل يرغبون في توسيع قاعدة معلوماتهم.

إن شيوع استخدام البريد الإلكتروني وشبكة الإنترنت حديثاً قد يشكل قوة دافعة جديدة.

استخدم مفهوم اللامركزية حديثاً جداً في ترجمة الأدوات الموجودة. ويرقى ذلك إلى حذف البنود الخاصة بثقافة معينة واستبدالها بأدوات محفزة أكثر ملاءمة. ذلك يعني في مصطلحات هذا الفصل خيار التكييف. مثال ذلك استخدم كورتيز وسميث (1979) هذه المقاربة لإعداد ترجمة إسبانية لاستبيان إنجليزي حول التثاقف.

يرتبط باللامركزية ما يدعى بالمقاربة المتلاقية Convergence approach (كامبل 1986). لنفترض أن عالماً نفسياً أميركياً وآخر من زمبابوي يهتمان بالمهارات التي يعتقد الأبوان أن من الضروري تطويرها عند الأطفال وبالسن التي يصبح الطفل بارعاً بها (تدعى النظريات الإثنية الأبوية: س. هاركنس وسوبر، 1992). قام كل من العاملين النفسيين بتطوير مقياس يتناسب مع البيئة الثقافية لكل منهما ومن ثم استخدمت هاتان الأداتان في كلا البلدين. في هذه الحالة تأخذ التشابهات والتباينات لتلك المعطيات أهمية خاصة. لا تعرف أي أمثلة تجريبية عن هذا الإجراء.

بعد جمع المعطيات من المجموعات المختلفة، هناك إجراءات لتحليل التكافؤ. تعالج بعض هذه الإجراءات تكافؤ البينة والمنهج في آن واحد. تبدو هذه الإجراءات جذابة بسبب قدرتها وسهولة تطبيقها. مثال ذلك يمكن سؤال الرواة المحليين أن يحكموا على صحة فهم الاختبار وملاءمته (الارشادات 2). وبشكل خاص يمكن للمجموعات ثنائية اللغة أن تقدم معلومات مفيدة حول كفاءة أحداث الاختبار (انظر، سيغي، الفصل الخامس من هذا الكتاب). وبالمثل يمكن إجراء مسوحات محلية لسكان اللغة المستهدفة تطرح فيها أسئلة حرة الاستجابة (الارشادات 6-3). يمكن الاختبار أن يستخدم أيضاً في اللغة المستهدفة بطريقة غير نظامية لطلب تفسير الأجوبة. تبين هذه الشروح ما إذا كانت الأسئلة قد فُسرت بالشكل الذي قصده مؤلف الاختبار.

إذا حصل تعديل جوهري في الاختبار أثناء عملية التكييف، يصبح من المهم بيان صدق بنية اختبار ما بعد تكييفه (الارشادات 20-10). مثال ذلك، تشونغ (189)، الذي كيف اختبار MMPI للاستخدام في الصين، درس قدرة القياس على التمييز بين الأشخاص الأسوياء والمرضى وأعد ملامح الشخصية بمجموعات مختلفة. دعمت كلاً من الطريقتين صدق البنية، إن تطبيق القوالب المتعدد السمات والمتعددة المناهج هي وسائط معقدة لمعالجة انحياز المنهج عند تكييف الاختبار يصعب الحصول عليها (كامبل وفيسك، 1959).

إن الأكثر شيوعاً من ذلك هو تطبيق تحليل العامل المؤكد وتحليل العامل الاستكشافي متبوع بتدوير الهدف (مثال، تايلوز وبوينس، 1991؛ واتكينس، 1989، ويندل وايواواكي ولينر، 1988) (مكاري وكوستا، 1997؛ بيد مونت وتشاي، 1997؛ شميدت وبه، 1992؛ فاند نبرغ وهاكستين، 1978).

إن الابتكار الحديث في تحليل العامل الاستكشافي هو تحليل العناصر المتزامن (كييرز، 1990؛ كييرز وتن بيرغ، 1989). خلافاً لما هو الحال في تحليل العامل الاستشكافي فإن جميع المعلومات تعالج مع بعضها، هناك مجموعة وحيدة من

العناصر الأساسية التي تُقيم عند سائر المجموعات. هذه العناصر الرئيسة هي بالتعريف متماثلة بين كل المجموعات؛ لذلك لا حاجة لتقويم الانسجام كما هو الحال في تحليل العامل الاستشكافي. إن نسبة التفاوت المبينة بوساطة العوامل الأساسية المشتركة تقارن مع تحليل العامل الأساسي في مجموعات منفصلة من المعطيات. يمكن إيجاد أمثلة عند زوكرمان، كوهلمان، ثورنكويست وكمبير (1991).

إن جميع الإرشادات الإدارية (18-13) واثنين من الإرشادات الوثائقية (22-19) يمكن اعتبارها مقترحات حول كيفية التقليل من انحياز المنهج. إن الطرق الفعالة لمعالجة انحياز المنهج تكمن في التدريب الشامل للذين يستخدمون الاختبار (ويتضمن ذلك التدرب على التواصل بين الثقافات فيما إذا كان المختبرون والمختَبرون من إثنيات مختلفة) وإعداد كتيبات متصلة خاصة بالاختبار تضمن تعليمات دقيقة. حتى في حالة تطبيق هذه المقترحات بدقة يبقى من المحتمل حدوث انحياز في المنهج يهدد صدق المقارنات عبر الثقافات. عندما تكون الفجوة الثقافية بين المجموعة الأصلية والمجموعة المستهدفة واسعة، فإن العينات ستختلف كثيراً في صفاتها ذات الصلة بالنتائج مثل التعليم، الحوافز، وأسلوب الإجابة، بحيث أن تقويم هذه الصفات يصبح الطريقة الوحيدة لضبط النتائج. إن الإجراءات الإحصائية مثل تحليل الخلافات المشتركة يمكن تطبيقها حينئذٍ لمعرفة إلى أي مدى يمكن لتباين الدرجات عبر الثقافات أن يكون ناجماً عن صفات خلفية العينات (الإرشادات 11).

هناك نوع آخر من الإجراءات لتحليل التكافؤ بين المعطيات وهو الاختبار وإعادة الاختبار أو الدراسة التدريبية التي تسمح بمقارنة نماذج المكتسبات عبر الثقفات. إن اختلاف نماذج المكتسبات يعطي دليلاً قوياً على سوء تكافؤ الدرجات. يمكن مشاهدة أمثلة على ذلك في نكيا، هوتو وبونت (1994). فقط طبق هؤلاء المؤلفون القوالب المعيارية لرافن ثلاث مرات على تلاميذ الصف السادس في فرنسا والكونغو. وقد وجدوا تحسناً معتدلاً في المرة الثانية ولم يجدوا أي تحسن في المرة

الثالثة في كل من المجموعتين عندما لم يكن هناك تحديد للزمن. مع ذلك عند تحديد الزمن أبدى أفراد المجموعتين تحسناً واضحاً في الدرجات في المرة الثانية كما أبدى التلاميذ من الكونغو تحسناً آخر في المرة الثالثة. إن هذه النتائج تعارض بشكل واضح قابلية مقارنة الدرجات في المرة الأولى من تطبيق الاختبار.

إن المجموعة الأخيرة من تقنيات تعزيز المصداقية هي التي تعالج انحياز البند (الإرشادات 9، 8، 7) وإن كلاً من إجراءات القياس السيكولوجي قد جرى ذكرها في مكان آخر من هذا الكتاب (انظر الفصل الأول والرابع).

الاستنتاج:

إن الشبهة بوجود انحياز ثقافي أو وجود بنية تجريبية على ذلك يعني أن الأداة (الاختبار) المقصودة لا يمكن اعتبارها بنية ذات أهمية أو نموذجاً عادلاً. إن أكثر الخطط المحافظة هي تلك التي تحاول أن تبرهن على أن الانحياز يمنع كل أشكال المقارنة. في اعتقادنا أن هذه الخطة هي الوحيدة الصحيحة عندما تتحدد البنيات حسب أسلوب ثقافة معينة. مع ذلك فإن الخطط البديلة ممكنه عندما تبدي البنيات تداخلاً عبر الثقافات وعندما يكون الانحياز غير ناجم عن البنية بحد ذاتها وإنما ينجم عن الطريقة التي يستعمل فيها اختبار معين. وقد بيّنا في هذا الفصل الفوارق المختلفة التي تساعد على معرفة الانحياز والنتائج السلبية التي تنتج عنه فيما يتعلق بتكافؤ درجات الاختبار عبر الثقافات والمقارنات التي تفيد في تجنب تلك النتائج.

إن تكييف الاختبار عبر الثقافات له جانب مفهومي وجانب مقياس متري، وقد أكدنا في هذا الفصل على الجانب الأول/ المفهومي.

إن استخدام الاختبارات عبر الثقافات وتفسير الاختلافات في نماذج الدرجات ومستوى الدرجات تواجه مآزق خطيرة، إلا أنه يمكن تجنب الكثير منها إذا كان معدو الاختبار ومستخدموه مدركين لوجودها.

*****

المراجع

Azuma, H., & Kashiwagi, K. (1987). Descriptors for an intelligent person: A Japanese study. *Japanese Psychological Research, 29,* 17–26.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments.* Thousand Oaks, CA: Sage.

Berry, J. W., & Bennett, J. A. (1991). Cree literacy. Cultural context and psychological consequences. *Cross-Cultural Psychology Monographs, no. 1.* Tilburg, Netherlands: Tilburg University Press.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology. Research and applications.* Cambridge, England: Cambridge University Press.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 389–444). Boston: Allyn & Bacon.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Newbury Park, CA: Sage.

Butcher, J. N. (Ed.). (1996). *International adaptations of the MMPI–2: Research and clinical applications.* Minneapolis: University of Minnesota.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Campbell, D. T. (1986). Science's social system of validity-enhancing collective believe change and the problems of the social sciences. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 108–135). Chicago: University of Chicago Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cheung, F. M. (1989). A review on the clinical applications of the Chinese MMPI. *Psychological Assessment, 3,* 230–237.

Cheung, F. M., & Leung, K. (1998). Indigenous personality measures: Chinese examples. *Journal of Cross-Cultural Psychology, 29,* 233–248.
Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Chang, J. P. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology, 27,* 181–199.
Church, T. A. (1987). Personality research in a non-Western setting: The Philippines. *Psychological Bulletin, 102,* 272– 292.
Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.
Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61–75.
Cole, M. (1996). *Cultural psychology. A once and future discipline.* Cambridge, MA: Harvard University Press.
Cortese, M., & Smyth, P. (1979). A note on the translation to Spanish of a measure of acculturation. *Hispanic Journal of Behavioral Sciences, 1,* 65–68.
Cotter, P. R., Cohen, J., & Coulter, P. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly, 46,* 278–284.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.
Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.
Erickan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2,* 199–215.
Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models. Foundations, recent developments, and applications.* New York: Springer-Verlag.
Gass, S. M., & Varonis, E. M. (1991). Miscommunication in nonnative speaker discourse. In N. Coupland, H. Giles, & J. M. Wiemann (Eds.), *Miscommunication and problematic talk* (pp. 121–145). Newbury Park, CA: Sage.
Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.
Greenfield, P. M. (1997a). Culture as process: Empirical methods for cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 1, pp. 301–346). Boston: Allyn & Bacon.
Greenfield, P. M. (1997b). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52,* 1115–1124.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.
Hambleton, R. K., & Swaminathan H. (1985). *Item response theory: Principles and applications.* Dordrecht, Netherlands: Kluwer.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Harkness, J. (1998). Cross-cultural survey equivalence [Special issue]. *ZUMA Nachrichten* (no. 3). Mannheim, Germany: Zentrum für Umfragen, Methoden und Analysen.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. Ph. (Eds.). (2003). *Cross-cultural survey methods.* New York: Wiley.

Harkness, S., & Super, C. (1992). Parental ethnotheories in action. In I. E. Sigel, A. V. McGillicuddy-DeLisi, & J. J. Goodnow (Eds.), *Parental belief systems: The psychological consequences for children* (2nd ed., pp. 373–392). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (pp. 155–165). Hong Kong: Oxford University Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296–309.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kiers, H. A. L. (1990). *SCA: A program for simultaneous components analysis.* Groningen, Netherlands: IEC ProGamma.

Kiers, H. A. L., & ten Berge, J. M. F. (1989). Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices for all populations. *Psychometrika, 54,* 467–473.

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das State–Trait Angstinventar. Theoretische Grundlagen und Handanweisung* [The State–Trait Anxiety Inventory. Theoretical background and manual]. Weinheim, Germany: Beltz Test.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76.

Liu, H. C., Chou, P., Lin, K. N., Wang, S. J., Fuh, J. L., Lin, H. C., Liu, C. Y., Wu, G. S., Larson, E. B., White, L. R., Graves, A. B., & Teng, E. L. (1994). Assessing cognitive abilities and dementia in a predominantly illiterate population of older individuals in Kinmen. *Psychological Medicine, 24,* 763–770.

Lucio, E., Reyes-Lagunes, I., & Scott, R. L. (1994). MMPI-2 for Mexico: Translation and adaptation. *Journal of Personality Assessment, 63,* 105–116.

Marcoulides, G. A., & Schumacker, R. E. (1996). *Advanced structural equation modeling: Issues and techniques.* Mahwah, NJ: Lawrence Erlbaum Associates.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509–516.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–118.

Merenda, P. F. (1994). Cross-cultural testing: Borrowing from one culture and applying it to another. In L. L. Adler & U. P. Gielen (Eds.), *Cross-cultural topics in psychology* (pp. 53–58). Westport, CT: Praeger/Greenwood.

Miller, J. G. (1997). Theoretical issues in cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 1, pp. 85–128). Boston: Allyn & Bacon.

Millsap, R. J., & Everson, H. T. (1993). Methodological review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.

Nkaya, H. N., Huteau, M, & Bonnet, J. (1994). Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills, 78*, 503–510.

Piedmont, R. L., & Chae, J-H. (1997). Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology, 28*, 131–155.

Poortinga, Y. H. (1975). Limitations on intercultural comparison of psychological data. *Nederlands Tijdschrift voor de Psychologie, 30*, 23–39.

Poortinga, Y. H., & van der Flier, H. (1988). The meaning of item bias in ability tests. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 166–183). Cambridge, England: Cambridge University Press.

Reese, S. D., Danielson, W. A., Shoemaker, P. J., Chang, T., & Hsu, H.-L. (1986). Ethnicity-of-interviewer effects among Mexican-Americans and Anglos. *Public Opinion Quarterly, 50*, 563–572.

Schmidt, S. M., & Yeh, R. (1992). The structure of leader influence: A cross-national comparison. *Journal of Cross-Cultural Psychology, 23*, 251–264.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard University Press.

Segall, M. H., Dasen, P. R., Berry, J. W., & Poortinga, Y. H. (1990). *Human behavior in global perspective. An introduction to cross-cultural psychology*. New York: Pergamon.

Serpell, R. (1979). How specific are perceptual skills? *British Journal of Psychology, 70*, 365–380.

Serpell, R. (1993). *The significance of schooling. Life-journeys in an African society*. Cambridge, England: Cambridge University Press.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–375.

Singer, E., & Presser, S. (1989). The interviewer. In E. Singer & S. Presser (Eds.), *Survey research methods* (pp. 245–246). Chicago: University of Chicago Press.

Sinha, D. (1997). Indigenizing psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 1, pp. 131–169). Boston: Allyn & Bacon.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–166.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State–Trait Anxiety Inventory ("Self-Evaluation Questionnaire")*. Palo Alto, CA: Consulting Psychologists Press.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*, 37–55.

Taylor, T. R., & Boeyens, J. C. (1991). The comparability of the scores of Blacks and Whites on the South African Personality Questionnaire: An exploratory study. *South African Journal of Psychology, 21*, 1–11.

Triandis, H. C. (1978). Some universals of social behavior. *Personality and Social Psychology Bulletin, 4*, 1–16.

Vallerand, R. J. (1989). Vers une methodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française [Toward a methodology for the transcultural validation of psychological questionnaires: Implications for research in the French language]. *Canadian Psychology, 30,* 662–680.

van de Vijver, F. J. R. (2003). Test adaptation/translation methods. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 960–964). Thousand Oaks, CA: Sage.

van de Vijver, F. J. R. & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 1, pp. 257–300). Boston: Allyn & Bacon.

van de Vijver, F. J. R., & Leung, K. (1997b). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29–37.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263–280.

van der Linden, W. J., &. Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Vandenberg, S. G., & Hakstian, A. R. (1978). Cultural influences on cognition: A reanalysis of Vernon's data. *International Journal of Psychology, 13,* 251–279.

Wagner, D. A. (1981). Culture and memory development. In H. C. Triandis & A. Heron (Eds.), *Handbook of cross-cultural psychology* (Vol. 4, pp. 187–232). Boston: Allyn & Bacon.

Watkins, D. (1989). The role of confirmatory factor analysis in cross-cultural research. *International Journal of Psychology, 24,* 685–701.

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of cultural anthropology* (pp. 398–419). New York: American Museum of Natural History.

Windle, M., Iwawaki, S., & Lerner, R. M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. *International Journal of Psychology, 23,* 547–567.

Zuckerman, M., Kuhlman, D. M., Thornquist, M., & Kiers, H. A. L. (1991). Five (or three) robust questionnaire scale factors of personality without culture. *Personality and Individual Differences, 12,* 929–941.

3

موضوعات أخلاقية منتقاة ذات علاقة بتكييف الاختبار

توماس أوكلاند
جامعة فلوريدا

يكتسب علم النفس أبعاداً عالمية (مايز، روبن، سابورين وواكر، 1996؛ روزينزيجر، 1999). تتوسع المنحة الدراسية، التي اعتمدت مرة وبشكل رئيس على مساهمات من أشخاص في أوروبا الغربية وشمال أمريكا، تتوسع لتشمل البحث العلمي وأشكالاً أخرى من المنح من علماء النفس في بلدان أخرى - في إفريقيا وآسيا المتحلقة بالمحيط الهادي وأوربا الشرقية وأمريكا الجنوبية. فوق ذلك تنمو تطبيقاته وتقنيته على نطاق واسع في بلدان عديدة.

تدرك الإسهامات المحتملة لعلم النفس في أهداف اجتماعية مهمة (مثلاً، التحصيل العالي التربوي، تطبيقات إدارية وصناعية مؤثرة وفعالة) وفي حلِّ موضوعات اجتماعية مربكة (على سبيل المثال: المرض العقلي، الإعاقة الشديدة، الفهم العنصري العرقي، وضبط السكان).

يمكن أن يُشاهد الانتشار العالمي لعلم النفس بوضوح تام في الاستخدام العالمي للاختبارات وأشكال أخرى من أدوات جمع المعطيات. يساعد استخدامها تحقيق حاجات متنوعة: تسهيل البحث، وصف السلوك، تمييز الموهبة، توثيق تحصيل المعرفة وقدرات ومهارات أخرى، تحسين الانتقاء التربوي والمهني، تشخيص

الاضطرابات، ومراقبة التغيير. إن هذه الحاجات عالمية وتحث على اتخاذ قرارات من دول عديدة لتطوير، وبطرق أخرى، لاكتساب تقنية اختبار تساعد على تحقيق تلك الحاجات.

التطور المبكر للاختبار

كانت الصين البلد الأول في استخدام الاختبارات بشكل واسع، وبعد أن طوّرت طرائق نمطية لتقييم الكفاءات المتعلقة بعمل الموظفين منذ أكثر من 3000 سنة (زانك، 1988). وقد حدث تطور واستخدام الاختبار في أوربا وشمال أمريكا بعد ميلاد علم النفس في الجزء الأخير من القرن التاسع عشر داخل أوربا الغربية.

الوضع الحالي لاستخدام الاختبار

لا تزال المعلومات حول الدرجة التي يجري فيها استخدام الاختبارات في أكثر من 200 بلد في العالم غير تامة. وتبقى الأبحاث العلمية على الاختبارات المستخدمة مع الراشدين غير منشورة (بارترام، كوين، 1998) وكذلك مناقشات الأنماط الإقليمية (بارترام، كوين، 1998، مونيز، بريتور، الميدا باراترام 1999). إن أبحاث عن اختبارات أجريت في نطاق عالمي على أطفال وشبيبة، هي أكثر واقعية، ذات أرضية عريضة وسهلة الوصول إليها (هيو وأولاكند، 1991: أوكلاند وهامبلتون 1995: أوكلاند وهيو، 1991، 1992، 1993).

لإعطاء تسلسل تاريخي، فقد جرى التعرف على 455 عنوان اختبارات تمّ استخدامها لتقييم أطفال وشبيبة في بلد واحد أو أكثر من ضمن 44 بلداً جرى مسحها، إن الاختبارات المستخدمة في العديد من تلك البلدان تم تطويرها بصورة نموذجية في داخلها وجرى الحصول عليها من الولايات المتحدة وبريطانيا أو فرنسا. وقد تمت ترجمة بعض الاختبارات إلى لغة البلد المعني: ولم يتم ذلك لبعضها الآخر. تبين أن توفر النماذج الوطنية سوية مع تقديرات للجدارة والصدق هو أقل تردداً في اختبارات مكيفة منه في الاختبارات التي تم تطويرها محلياً (أوكلاند وهيو، 1991).

غالباً ما تعترض شروط متنوعة محاولات تطوير الاختبارات في كل بلد. وتشمل تلك عدم توفر الاختصاصيين في تطوير الاختبار، توجيهات اجتماعية وسياسية تُقلّل من متانة صلة الفرق الفردي بمواقف المساواة البراقة، واعتماد علم النفس على النظرية أكثر منه على المبادرات التجريبية. بالإضافة إلى ذلك، فإن قليلاً من سكان بلدان عديدة بالإضافة إلى الفشل في حماية حق التأليف للاختبارات يُقلّل من المردود المادي المشروع للاستثمار المرافق لتطوير الاختبارات.

يجري استخدام الاختبارات المكيّفة تحت واحد من الشروط الثلاثة التالية: للاستخدام في بلدان غير تلك التي جرى تطويرها فيها؛ للاستخدام في البلدان التي جرى تطويرها فيها مع تكييفها لتستخدم مع أشخاص يختلفون في اللغة، الثقافة، أو صفات مهمة أخرى؛ وللاستخدام في بلدين أو أكثر حيث تجري تطبيقات على قوميات متعددة.

إن الشرط الأول عام (مثلاً تكييف الاختبارات للاستخدام في بلدان غير التي جرى تطويرها فيها).

على سبيل المثال: وجد أوكلاند وهيد (1993) أن اختبارات عديدة تم تطويرها بالأصل بالولايات المتحدة، وبريطانيا، وفرنسا، وتمَّ تكييّفها للاستخدام في بلدان أخرى (مثلاً تكييف مقياس الذكاء الذي تم تطويره في الولايات المتحدة لأجل الاستخدام في الكويت).

ويصبح الشرط الثاني أكثر عمومية (مثلاً، تكييف اختبارات لتحسين صدقها من أجل استخدامها في بلدان أخرى) كمهاجرين فارين من أراضي أوطانهم المُخربة بالحرب، زيادات الهجرة الناجمة عن تغييرات اقتصادية وسياسية، تقسيمات حدودية تقليدية يتم محوها من قبل حلفاء سياسيين جدد (مثلاً، الاتحاد الأوربي) وشروط أخرى تُسهل تدفق الأشخاص إلى داخل قطاعات جغرافية وسياسية جديدة. على سبيل المثال، عدد اللغات الأولى المنطوقة من قبل التلاميذ في بعض

مناطق المدارس العامة يزيد على 150 . في داخل الولايات المتحدة يطالب القانون المدارس أن تقيم الأطفال حسب لغتهم الغالبة عند تصنيفهم في صفوف خاصة. وبذلك يمكن أن تساعد وفرة الاختبارات المكيفة لاستخدامها مع أشخاص يعتمدون على لغات أجنبية يمكن أن تساعد أقسام المدرسة على أن تذعن للقوانين بقدر ما تُسهل التقييم.

أثار الاعتراف بأننا نعيش في مجتمع عالمي وكوني اهتماماً معتبراً في أنشطة الاختبار المتصل بالقوميات المختلطة، وعلى وجه الخصوص في الصناعة والتربية. على سبيل المثال، طورّت شركات متعددة الجنسيات تطبيقات توظيف ذاتية تعتمد على اختبارات مكيّفة. بالإضافة إلى ذلك، عند إبداء الرغبة لتطوير برامج تربوية على مستوى عالمي، تشارك دول عديدة في دراسات عالمية في الرياضيات، العلوم، وتحصيل القراءة؛ تعتمد تلك الدراسات بشكل قوي على الاختبارات المكيّفة. كما ستتم الإشارة لاحقاً، تعتمد مؤسسات متنوعة حكومية وغير حكومية، ومتعددة الجنسيات على معلومات من دراسات القوميات المتداخلة عندما ترسم سياسة وتؤسس تطبيقات. سوف تؤثر نوعية الاختبارات المكيفة المستخدمة لجمع المعطيات على صدق تلك المعلومات وبالتالي فائدتها.

إحدى عشرة مجموعة معنية بالاختيار

تميل إحدى عشرة مجموعة من الأشخاص أو أكثر إلى أن تكون منغمسة باستخدام الاختبارات المكيفة أو متأثرة بها. (انظر الجدول 103) وهي تشمل ما يلي:

1- يطور مؤلفو الاختبار اختباراً يجري تكييفه لاحقاً بتفويض منهم. يمكن أن يجري تعديل الاختبار كاملاً إما مع أو دون اشتراك المؤلفين ومساعدتهم. على سبيل المثال: جرى تكييف البيان المفصل للقلق ذي السمة الرسمية إلى أكثر من 60 لغة ولهجات محلية. وقد ساعد مؤلفها الدكتور تشارلز سبيلبيرجر Charles spiel-berger في حوالي 25٪ إلى 30٪ من تلك التكييفات.

الجدول 1-3

إحدى عشرة فئة من أشخاص شاركوا في ترجمة اختبارات أو استخدامهم

1- مؤلفو الاختبار الذين طوّروا اختباراً جرى تكييفه لاحقاً بتفويض منهم.

2- شركات نشرت ووزعت اختبارات جرى تكييفها لاحقاً بتفويض منهم.

3- مؤلفو الاختبار الذين طوّروا اختباراً جرى تكييفه لاحقاً دون تفويض منهم أو من الناشر.

4- مختصون جرى استخدامهم لتطوير اختبارات للاستخدام في لغات وبلاد متعددة.

5- مختصون يساعدون في الاختبارات المكيّفة.

6- مختصون يدربون آخرين على استعمال طرق تكييف الاختبار.

7- أشخاص أو منظمات تحتاج إلى معلومات اختبار لاتخاذ قرارات.

8- مختصون ينتقون اختبارات مكيّفة ويستخدمونها لاكتساب المعلومات.

9- أطراف ثالثة (مثلاً: المدراء) يستخدمون البيانات من اختبارات مكيفة.

10- أشخاص يجري اختبارهم باختبارات مكيفة وقرارات يتم اتخذها بصددهم.

11- مستهلكون لمعلومات جرى الحصول عليها من دراسات عبر القوميات تستخدم اختبارات مكيفة.

أخبر (سبيلبرغر، اتصالات شخصية، 15 أيلول 1999) إن معظم التعديلات غير مُرخصة.

2- نشرت شركات الاختبار اختبارات ووزعتها وجرى تكييفها فيما بعد بترخيص منها. يمكن أن تكون ترجمة الاختبار تامة إما مع انغماس الشركة ومساعدتها أو دونها، وجرى إعطاء مثلين: أعطى الدكتور ريتشارد وودكوك، المؤلف الرئيس لـ وودكوك - جونسون البطارية المنقحة للتربية النفسية (WJ-R)، وناشرها

Riverside Press، أعطى ترخيصاً لتكييف WJR لأجل استخدامها في جمهوريات السلوفاك والتشيك، هنغاريا ولاتفيا. وهم أيضاً يساعدون في تلك التكييفات، وقد رخّصت مؤسسة تقييم هاركورت 26 تعديلاً رسمياً من ثلاث تراجم لمقياس فكسلر لذكاء الأطفال WISC، WISC، WISC-R، WISC-III، ومنحت الإذن لترجمة واستخدام WISC في لغات إضافية على أساس حالة بحالة للاستخدام في دراسات محددة فقط. ينبغي إتلاف النسخ الباقية عند إتمام الدراسة. لم تساعد شركة تييم هاركورت في تلك التكييفات (ل. مورفي، اتصالات شخصية، آب 1999).

3- يطور مؤلفو الاختبار اختباراً يجري تكيفه لاحقاً دون تفويض منهم أو من الناشر. على سبيل المثال، تُؤسس الطبعات المتنوعة لـ WISC المقياس المعتمد لذكاء الأطفال الأكثر تكراراً للاستخدام بصورة فردية (أوكلاند وهيو 1992). بالرغم من أن مؤسسة هاركورت للتقييم قد رخصت لـ WISC 26 تكييفاً رسمياً، تمَّ تسجيل استخدامه في 34 بلداً على الأقل. وبالتالي يوجد عدة طبعات غير مرخصة.

4- يتم توظيف مختصين مدربين في تطوير الاختبار لأجل تطوير اختبارات بغية استخدامها في لغات وبلدان متعددة. على سبيل بالمثال، تُنتج هيئة البرنامج المهني الموثق لمايكروسوفت سنوياً 45 اختباراً وأكثر بـ 16 لغة لأجل الاستخدام في 75 بلداً للمساعدة على توثيق كفاءات ملايين الأشخاص الذين يستخدمون منتجاتها (فيترجيرالد وورد، 1998).

5- يساعد المختصون في تكييف الاختبارات. ربما يشمل عملهم ترجمة لغة اختبار، تطوير وتنقيح مواد، تجميع معطيات ينجم عنها تأسيس أنماط وطنية جديدة، مثل ملامح سيكولوجية أخرى للاختبار (مثلاً، تقديرات الجدارة والصدق). ربما تكون هذا المبادرة شخصية أو يحدث بناء على طلب طرف آخر (مثلاً: موزع اختبار يكون بحاجة إلى اختبار مكيّف).

6- إن أولئك الذين يدربون آخرين على طرق تكييف اختبار يشكلون أيضاً عنصراً مهماً. يجب أن يتأكدوا أن طبيعة تحضير الطلاب العلمية والمهنية هي متداولة وعميقة.

7- يؤثر استخدام اختبارات مكيّفة على خمس مجموعات للمستهلك. الأول يتألف من أشخاص أو منظمات تحتاج إلى معلومات اختبار للتوصل إلى قرارات. يتلو ذلك بعض الأمثلة. يحتاج عالم نفسى نرويجي إلى تحديد القدرات العقلية لمراهق من أوساط إفريقية. تحتاج شركات إسترالية متعددة الجنسيات إلى استئجار مدراء وسطاء إضافيين في مكتبها في آسيا. تحتاج لجنة القبول للجامعة في كندا إلى أخذ قرار حول طلب من طالب متخرج طموح من أواسط أمريكا. يبحث الاتحاد الأوربي للتأكد من معايير مهنية مقارنة للأطباء عبر أرجاء أوربا عن طريقة استخدام مقياس عام لكفاءة التطبيق الطبي. ربما تثير تلك الحاجة وأخرى غيرها من أشخاص أو منظمات حاجة لتكييف اختبار ما.

8- تتألف مجموعة المستهلك الثانية من محترفين ينتقون اختبارات مكيّفة ويستخدمونها لاكتساب المعلومات. يدير المعلمون، المستشارون، الممرضات، الأطباء، المديرون، علماء النفس ومحترفون آخرون متنوعون (أوكلاند وهيو 1991) يدير هؤلاء بشكل روتيني الاختبارات ويضعون الدرجات النهائية لها كما يفسرون نتائجها ويكتبونها في تقرير وذلك للمساعدة في صنع القرار.

9- غالباً ما يختبر المختصون المحترمون في استخدام الاختبار أشخاصاً (مثال الأقسام الأولى) لمساعدة آخرين (مثال الأقسام الثالثة) في صُنع قرارات اختبارية وسارية المفعول حول أشخاص جرى اختبارهم. على سبيل المثال: ربما يشمل أقصى عدد من متلقي معلومات الاختبار، محترفي القسم الثالث المتنوعين (مثل موظفي الموارد البشرية، مديرين، مربين، أطباء، وقضاة) الذين ربما يطلبون إجراء الاختبار على شخص أو آخر، وهم يشكلون مجموعة المستهلك الثالثة.

10- تتألف مجموعة المستهلك الرابعة من أشخاص جرى اختبارهم بواسطة اختبارات مكيّفة بعد أن يجري اتخاذ القرارات حولهم، وتؤثر الدرجة التي يعرض المحترفون فيها أنماطاً سلوكية أخلاقية. تؤثر بشدة على صدق القرارات المتعلقة بالاختبار والمُتخذة حول أولئك الذين جرى اختبارهم.

11- تشكل مجموعة المستهلك الخامسة الهيئات الحكومية وغير الحكومية، المؤسسات متعددة الجنسيات ومستهلكين آخرين لمعلومات تم الحصول عليها من دراسات لقوميات متداخلة أو ثقافات متداخلة، وهم يعتمدون على تلك المعلومات عند رسم سياسة أو ترويج تطبيقات. بالإضافة إلى ذلك، يدير عدد من علماء السلوك البحث في القوميات المتداخلة. سوف تؤثر نوعية الاختبارات المكيّفة المستخدمة في جمع المعطيات على الصدق وبالتالي على فائدة المعلومات التي يتلقونها.

إرشادات ومعايير الاختبار

الإرشادات العامة والمعايير:

يمكن أن توجه الإرشادات والمعايير الجهود لتعزيز تطبيقات ملائمة تحكم تطور الاختبارات المكيّفة واستخداماتها. يشمل الصف الأول وثائق تحمل عناوين موضوعات مهمة وواسعة تؤثر على تطور الاختبار واستخدامه في مناطق ثلاث: تلك التي تتوجه إلى الأبعاد الفنية وذات العلاقة بالمفاهيم لتطوير الاختبار واستخدامه (e.g. المعايير للاختبار النفسي والتربوي، المُعد من قبل جمعية البحث التربوي الأمريكي، جمعية علم النفس الأمريكية [APA]، والمجلس القومي للقياس في التربية، 1999)، تلك التي تُميّز المهارات المهنية والقدرات المحتاجة من قبل أولئك الذين يستخدمون الاختبارات (e.g. جمعية علم النفس البريطانية، 1998، 1999؛ أيد، مورلاند، روبرستون، بريموف وموست، 1988؛ لجنة الاختبار العالمية، 2000؛ اللجنة المشتركة حول تطبيقات الاختبار، 1988) وتلك التي تتوجه إلى الموضوعات الأخلاقية.

بالرجوع إلى الموضوعات الأخلاقية المقترنة بتطوير الاختبار واستخدامه، حدّدت بعض الجمعيات المهنية بشكل جيد المعايير الأخلاقية التي تتوجه إلى طائفة واسعة من الموضوعات (APA, 2000؛ الجمعية النفسية البريطانية، (1998b بعض تلك الموضوعات تركّز فقط على تطوير الاختبار واستخدامه. دساتير ووثائق أخرى تتوجه بصورة مباشرة أكبر إلى موضوعات اختبار (e.g.) الجمعية النفسية الكندية، 1987؛ اللجنة المشتركة لتطبيقات الاختبار، 1998؛ كندال، جينكنسون، ديلموس وكلانسي، 1977؛ المجلس القومي للقياس في التربية، 1995؛ كوين، 1997؛ لندس، 1996). سوف تستفيد الجهود لترويج تطبيقات اختبار صلدة مع مقاييس مكيّفة، سوف تستفيد من فتحة تعليمية في تلك المجالات الثلاث.

إرشادات ومعايير أكثر تركيزاً على اختبارات مكيّفة

بالإضافة إلى ذلك نحن بحاجة إلى صف ثان من الوثائق. تلك التي تتوجه إلى موضوعات أكثر تركيزاً على تكييف الاختبارات واستخداماتها. مرة ثانية، نحن بحاجة إلى وثائق تتوجه إلى معايير فنية، مهارات وقدرات مهنية، وأخلاقيات.

المعايير الفنية

يتم صنع التقدم في تطوير توصيات فنية وذهنية للمساعدة على توجيه التكييفات المطبقة عالمياً (هامبلتون، 1994؛ 2001؛ انظر كذلك الفصل الأول، هذا المجلد). إن تلك الارشادات تعد حاسمه هي دقيقة لتأسيس تطبيقات متينة بغرض تكييف الاختبارات وتقرير مساواة الدرجات النهائية فيها. يجري إخضاع تلك التوصيات الفنية وأخرى غيرها إلى مراجعات مستمرة.

مهارات وقدرات مهنية

إن الارشادات والمعايير التي تناقش النوعيات المهنية لأولئك الذين يستخدمون اختبارات مكيفة غير متوفرة. على أية حال، إن عدداً من مقومات الإرشادات العالمية لاستخدام الاختبار (لجنة الاختبار العالمية، 2000) ومصادر أخرى (مثال اللجنة

المشتركة لتطبيقات الاختبار، 1988؛ ايد، وآل، 1988) تكون ملائمة لتكيفات الاختبار.

الأخلاق

لم يكن بالإمكان إيجاد كتابات علمية تدور حول موضوعات أخلاقية متعلقة بتطوير واستخدام اختبارات مكيّفة. لا توجد معايير عالمية أو إرشادات تتوجه إلى تلك الموضوعات. لذلك على المرء في الوقت الحاضر أن يعتمد على وثائق أخرى تتوجه بشكل أوسع إلى الموضوعات الأخلاقية.

على سبيل المثال، تناقش الإرشادات العامة لاستخدام الاختبار (بارترام، 2001؛ لجنة الاختبار العالمية، 2000) خمس موضوعات أخلاقية مهمة وواسعة الانتشار:

(أ) الحاجة إلى العمل بطريقة أخلاقية ومهنية، (ب) التأكيد أنه على أولئك الذين يستخدمون الاختبارات أن يتوخوا الكفاءات، (ج) أن تكن مسؤولاً لاستخدام الاختبار، (د) التأكيد على أن مواد الاختبار آمنة، و(هـ) التأكيد على أن نتائج الاختبار سرية. وتشمل كشوفات أخلاقية من هيئات أخرى (مثال: APA 2002؛ الجمعية النفسية البريطانية، 1998b؛ اللجنة المشتركة لتطبيقات الاختبار، 1988؛ كندال وآل، 1997؛ المجلس القومي للقياس في التربية، 1995) تشمل أيضاً: قضايا تتصل باختبارات مكيّفة وينبغي الاستشارة بصددها.

الغرض من هذا الفصل

يمكن أن يشجع نقاش حول موضوعات أخلاقية أكثر تحديداً لترجمة الاختبارات تطبيقاً مهنياً وفنياً صلداً في هذا المجال الناشئ. هكذا يكون الغرض الأول من هذا الفصل هو مراجعة مبادئ أخلاقية منتقاة ومعايير من أحد المبادئ الأخلاقية الراسخة جيداً (.i..e المبادئ الأخلاقية لعلماء النفس ومبادئ السلوك، أشير إليها فيما بعد بالمبادئ الأخلاقية، منشورات الجمعية النفسية الأمريكية، 2002) في ضوء تطبيقات متنوعة يمكن أن تقترن باختبارات مكيّفة واستخداماتها.

تجرى فهرسة أنماط سلوك إحدى عشرة مجموعة موصوفة سابقاً ومساهمة في اختبارات تكيّف واستخدام معلومات منها، تجري فهرستها في ضوء 24 معياراً أخلاقياً. يجري صنع المرجع لإحدى عشرة جمعية المذكورة آنفاً في محاولة جاهدة لاقتراح تلك المجموعات التي ربما تكون الأكثر حساسية لكل من الخمس والعشرين معياراً.

مبادئ أخلاقية عامة

ترتكز الدساتير المتوجهة إلى أنماط السلوك الأخلاقي تماماً على مبادئ عامة. على سبيل المثال: يمكن أن يشكل المبدأ العام "أن لا تسبب أي أذى" الحجر الأساسي للمبدأ الأخلاقي الذي ينفذ إلى كل المبادئ الأخلاقية المهنية (كووشر وكيث سبيخل، 1998). يرتكز الدستور الأخلاقي على خمسة مبادئ حساسة. يتم ذكرها أدناه وتستخدم كأساس لأجل فهم الخمس والعشرين معياراً أخلاقياً التالية:

الإحسان وعدم الإضرار: يحاول علماء النفس جاهدين ليفيدوا أولئك الذين يعملون معهم والذين يخدمونهم. إنهم يتطلعون إلى حماية مصلحة وحقوق أولئك الذين يتفاعلون معهم. إنهم يلتزمون الحيطة والحذر تجاه العوامل الشخصية، المالية، الاجتماعية، التنظيمية، أو السياسية، التي ربما تؤدي إلى سوء استخدام تأثيرهم.

الإخلاص والمسؤولية: يؤسس علماء النفس علاقات ثقة ويدعمونها مع أولئك الذين يعملون معهم. إنهم مدركون لمسؤولياتهم العلمية والمهنية تجاه المجتمع والجاليات الخاصة التي تعمل فيه. إنهم يؤيدون معايير سلوك مهنية، يوضحون دورهم والتزاماتهم المهنية، يتحملون مسؤولية مناسبة لسلوكهم، ويتطلعون إلى معالجة تضارب المصالح الذي ربما يؤدي إلى الاستغلال أو الأذى.

الاستقامة: يتطلع علماء النفس إلى تعزيز الدقة، الأمانة، والصدق في العلم، التعليم، وتطبيق علم النفس. إنهم لا يسرقون ولا يغشون أو يشاركون في خداع أو احتيال، أو يُحرّفون الحقائق على نطاق عالمي. إنهم يحاولون جاهدين المحافظة على وعودهم وتجنب التزامات غير حكيمة أو واضحة.

العدالة: يعترف علماء النفس أن الإنصاف والعدالة يمكنان كل الأشخاص أن ويستفيدوا من إسهامات علم النفس وإلى نوعية مساوية في العمليات والإجراءات، والخدمات التي يقدمها علم النفس. يتخذ علماء النفس احتياطاتهم على أن لا تؤدي انحيازاتهم المحتملة، حواجز الكفاءة وحدود خبرتهم إلى التغاضي عن تطبيقات غير عادلة.

احترام حقوق الناس وكرامتهم: يحترم علماء النفس كرامة وأحقية كل الناس، وحقوق الأفراد للخصوصية، والسرّية، والقرار الشخصي. يدرك علماء النفس، خلال عملهم، احترام الفوارق الفردية والثقافية والوظيفية، بما فيها تلك المبنية على العمى، الجنس، الهوية الجنسية، العنصر، العرق، الثقافة، الأصل، القيم، الديانة، التوجه الجنسي، العجز، اللغة، والوضع الاجتماعي الاقتصادي.

خمس وعشرون معياراً أخلاقياً

إن الخمسة والعشرين معياراً التالية هي بين 89 معياراً تؤلف الدستور الأخلاقي لـ .APA مقابل خمسة المبادئ الموصوفة سابقا والتي هي غير ملزمة، تصف المعايير أنماط السلوك التي من المتوقع أن يبديها الأعضاء وهي ملزمة؛ ربما تقود الانتهاكات إلى عقوبات، يُقصد بالأمثلة المستعملة هنا تجديد مجالات استعمالات محتملة لتطبيقات مقترنة باختبارات مكيّفة واستخداماتها. لم يجر اشتقاق الأمثلة من مسح حادثة وليس المقصود منها أن تستنفد كل الموضوعات المحتملة والمشكلات التي ربما ترافق هذا العمل.

2.01 حدود الكفاءة:

(أ) يقدم علماء النفس خدمات، يعلمون، يجرون بحثاً مع السكان وفي مناطق فقط ضمن حدود كفاءتهم، المرتكزة على تعليمهم، التدريب، الخبرة الموجهة، الاستشارة، الدراسة، أو الخبرة المهنية.

مثال: إن العلماء الذين يساعدون في الاختبارات المكيّفة يستخدمونها في العمل التطبيقي، وتعمل الاختبارات المصممة غالباً في مياه مجهولة. إن الإرشادات التي تتحكم بتجربة الإختبار هي جديدة ومتطورة. الكتابات حول هذه المسألة ضئيلة. على نطاق العالم يوجد برامج قليلة للمساعدة في إعداد المهنيين في هذا المجال المهم والمتخصص. وهكذا تكون معظم المعرفة المخصّصة لاختبار التكيفات واستخداماتها مكتسبة ذاتياً.

ينبغي للمعرفة المكتسبة ذاتياً أن تُدعّم بواسطة التربية الأساسية، والتدريب والخبرات المُراقبة عندما يكون ذلك ممكناً. وفوق ذلك، يمكن أن ينجم عن العمل المُنجز كفريق حيث يرعى وينصح الأعضاء بعضهم بعضاً ويشرف كل منهم على الآخر، ربما يُنتج ذلك مستويات عالية أكثر من أن يتم إنجاز العمل بصورة فردية.

إن معرفة الكتابات حول الطرائق التي ينبغي استخدامها عند تكييف الاختبارات، وتشمل الإرشادات لتكييفاتهم (هامبلتون 1994، 2001؛ انظر أيضاً الفصل الأول، هذا المجلد)، هي مُلزمة لهذا العمل. على أية حال، لم يتم معرفة هذه الكتابات على نطاق واسع وبالتالي لم يتم تطبيقها على نطاق واسع. فوق كل ذلك، تخضع الإرشادات إلى المراجعة عندما يُطوّر المهنيون كلاً من النظرية والتقنية لتحسين تكييفات الاختبار (مثال: هامبلتون، 2001). مجموعات 1 - 9 في الجدول 3-1 ربما تتأثر بهذا المعيار.

ينبغي على أولئك الذين يستخدمون الاختبارات المكيفة، أن يكونوا حذرين بصورة خاصة نظراً لحداثة هذا الميدان. ينبغي عدم الادعاء أن اختباراً مكيفاً يشبه الاختبار الأصلي.

2.03 المحافظة على الكفاءة

يأخذ علماء النفس على عاتقهم الجهود المتقدمة لتطوير كفاءتهم وصيانتها.

مثال: ينبغي أن يكرس الأشخاص العاملون في الاختبارات المكيفة واستخداماتها لهذا العمل جزءاً أساسياً من حياتهم المهنية لكي يصبحوا ضليعين في

العديد من جوانبه المعقدة. ومن المتوقع، أن تتطلب الكفاءة معرفة باللغات، وعلم النفس المعرفي والنموّ، والفروق الفردية، وعلم الإنسان الاجتماعي والثقافي، وعلم الاجتماع، والقياس السيكولوجي والإحصاء، ومعرفة بالمواقع التي سيتم استخدام الاختبارات فيها. إضافة إلى ذلك يتطلب صيانة ونمو المعرفة، والتطبيقات في أي ميدان ناشئ مؤسسة فرعية تتضمن قادة منطقة وعلماء، وحضور مؤتمرات عالمية ووطنية، وطرائق مساعدة على سرعة الحركة أو العمل وذلك لاكتساب المعرفة الفنية للحالة. يمكن أن تكون المعلومات المقدمة من المجلات والكتب، رغم كونها مفيدة، مدعمة في الوقت الذي نشرت فيه، ربما تكون مجموعات 1-8 (الجدول 1-3) متأثرة بهذا المعيار.

2.04 أسس الأحكام المهنية والعلمية:

يرتكز عمل علماء النفس على معرفة تامة علمية ومهنية بمجال التخصص.

مثال: يجري تأسيس العديد من العناصر المهمة لتطوير نوعية الاختبارات المكيّفة واستخدامها جيداً بصورة متينة وطويلة الأجل. على سبيل المثال: تملك الطرائق الكمية والمقترنة بتطوير الاختبار، وتشمل تلك التي تؤسس أنماطاً وتقدر الجدارة والصدق تقاليد طويلة في علم النفس الغربي وتشكل بضع الأعمدة الأكثر رسوخاً في علم النفس (مثال: أناستاسي وأوربينا، 1997؛ امبرتسون وهرشبرغر، 1999؛ هالادينا، 1999؛ ماكدونالد، 1999). تسهم الإنجازات الفكرية والنظرية في تطوير الاختبار خلال العقدين الماضيين أيضاً في قوتنا المؤسساتية (مثال: بايرن، 1998، لوهن، 1998؛ شوماكر وماركولدز، 1998). وبذلك يرتكز الأساس للأحكام العلمية على أرضية صلبة.

بالإضافة إلى ذلك، يجد الحكم المهني في استخدام الاختبارات أيضاً دعماً في تقاليد مئة عام من استخدام الاختبارات لاتخاذ قرارات عملية متعلقة بالأفراد والجماعات، في برامج مهنية متنوعة للخريجين، تُعدّ مهنيين لاستخدام الاختبارات، وفي الكتابات العلمية الفنية حول استخدام الاختبار (مثال: شاتلر، 1988).

تعطي هذه المعرفة شرعية ثرية. ويكون أساس المعرفة هذا حيوياً للمهنيين العاملين في الاختبارات المكيّفة. وهذه المعرفة دقيقة بصورة خاصة لعمل أولئك الذين يُصممون ويُطورون الاختبارات المكيّفة وفي طرق أخرى يساعدون في تكييفها. بالرغم من كون هذه المعرفة أقل دقة فإنها لا تزال مفيدة لأولئك الذين يستخدمون الإختبارات المكيّفة كما هي بالنسبة لمستهكي النتائج من الدراسات التي تعتمد على الاختبارات المكيّفة. وإذا طرأ سؤال، ينبغي أن يكون المهنيون قادرين على الإشارة إلى الكتابات المهنية والعلمية كأساس لعملهم. ربما تتأثر كل المجموعات (الجدول 1-3) بهذا المعيار.

2.01 حدود الكفاءة:

(أ) حيث يتم ترسيخ المعرفة المهنية أو العلمية في نظام علم النفس ويكون فهم العوامل المقترنة بالعمر، والجنس، بالهوية الجنسية، بالعنصر، بالعرق، بالعجز، باللغة، وبالوضع الاجتماعي الاقتصادي، جوهرياً بغية التنفيذ الفعال لخدماتهم أو بحثهم العلمي، يملك علماء النفس أو يحصلون على التدريب والخبرة والاستشارة، أو الإشراف الضروري لتعزيز الكفاءة في خدماتهم، أو أنهم يكوِّنون مراجع استشارية ملائمة.

مثال: أعطت دراسة الفرق الفردية في القدرات الذهنية وغيرها من صفات الشخصية دفعة إلى الأمام لتخصص علم النفس. ويبقى علم النفس ملتزماً بدراسة الفروق الفردية.

تمتد دراسة الفروق الفردية غالباً إلى دراسة فروق المجموعات. وقد كشف البحث التطبيقي في علم النفس أن الفروق الذهنية المهمة والمقبولة وصفات شخصية أخرى موجودة كوظيفة العمر، والجنس، والعنصر، والعرق، والوضع الاجتماعي الاقتصادي وصفات ديموغرافية أخرى. يتم كشف تلك الفروق غالباً من خلال معطيات الاختبار (هرنستبن وموري، 1994؛ جنسن، 1980).

على أية حال، لم يقبل بعض الناس اكتشافات البحث تلك وعوضاً عن ذلك يعتقدون أن فروق المجموعات تُؤدي إلى اختبارات منحازة وغير صالحة (مثال: ميرسر، 1973؛ اوكلاند، 1977؛ رينولدوز وبراون، 1984). يجري علم أولئك العاملين في الاختبارات المكيّفة واستخداماتها في بيئات ثقافية واجتماعية. ويسمع المرء صوت وجهات نظر ومواقف الرأي العام قوياً معبراً عن نقاط مهمة، ينبغي عدم تجاهلها.

ينبغي أن يكون أولئك العاملون في الاختبارات المكيّفة واستخداماتها حساسين لوجهات النظر والمواقف؛ لأن الاختبارات تتجه إلى أن تكون غير صادقة عندما يجري استخدامها مع جماعات مختلفة في بلد واحد وخاصة عندما يتم استخدامها مع قوميات متعددة. هناك حاجة لجهود للتأكيد أن للاختبارات المستخدمة مع مجموعات مختلفة مصداقية مقارنة، ويحتاج المرء إلى درجات نهائية متوازنة للتغلب على التوقعات السلبية (هامبلتون، 1994، 2001؛ انظر أيضاً إلى الفصل الأول، هذا المجلد).

ينبغي أن يتم استخدام الاختبارات المكيفة فقط بعد عرض صفات سيكولوجية مبنية على معطيات مكتسبة من السكان هدف البحث. إضافة إلى ذلك، يتخذ المحترفون الذين يتبنون مواقف منحازة تجاه مجموعة واحدة أو أكثر (مثال: العمر، الجنس، العنصر، العرق، الوضع الاجتماعي الاقتصادي) والتي ربما تؤثر على عملهم، يتخذون، خطوات للتغلب على انحيازاتهم أو ينسبون العمل إلى آخرين لا يكون لديهم تلك الانحيازات. ربما تكون المجموعات 1-5 و7-11 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

1.01 سوء استخدام عمل علماء النفس

إذا عَلِمَ علماء النفس عن سوء استخدام أو سوء تمثيل لعملهم يتخذون خطوات معقولة لتصحيح سوء الاستخدام أو التقليل منه أو من سوء تمثيله.

مثال: بمناسبة ما، يمكن أن يُطلب من المهنيين أن يعملوا في اختبارات مكيّفة أو استخدامها عندما يكون للعمل بصورة بادية للعيان أهلية مشكوك فيها. على

سبيل المثال، بالرغم من أن امتلاك المعرفة بأن عينة ما غير ممثلة في مجموعة السكان الهدف، ربما يُطلب من المحترفين استعمال مجموعة معطيات موجودة، ربما تكون حتى في نسخ عينات فرعية، وذلك في محاولة للحصول على أنماط أوسع وأكثر تمثيلاً أو لتأسيس صفات سيكولوجية أخرى. ربما يُطلب من علماء النفس التطبيقيين استخدام اختبار مكيّف يفتقد إلى صدق مبني على التجربة ووصفه كمشابه لأحد الاختبارات القياسية المستخدمة في الصناعة. ينبغي تجنب تلك الحالات.

ينبغي أن لا يجري استخدام اختبار مُكيّف يبدو أنه صادق، ومرتكز على معايير ذات صدق ظاهري، إذا افتقد البرهان التقني الذي يدعم صدقه للاستخدام مع المجموعة الهدف ولأجل أغراض محددة. لا يمارس علماء النفس أنشطة تُعرّض على الأرجح قدراتهم ومهاراتهم إلى سوء الاستخدام. من المفروض على المهنيين أن يتكلموا بصوت عال عندما تحدث انتهاكات أو سوء استخدامات وأن يتخذوا خطوات معقولة لتصحيحها والتقليل منها. ربما تكون المجموعات 6، 8 و9 (الجدول 3-1) الأكثر تأثراً بهذا المعيار.

3.05 علاقات متعددة:

(أ) تنشأ العلاقة المتعددة عندما يكون عالم النفس في دور مهني مع الشخص و(1) يكون في الوقت عينه في دور آخر مع الشخص نفسه، (2) في الوقت عينه يكون في علاقة مع شخص مرتبط بشدة أو له علاقة بالشخص الذي يكون لعالم النفس علاقة مهنية به، أو (3) وعود بإجراء علاقة في المستقبل مع الشخص أو بشخص مرتبط به بشدة أو له علاقة معه.

يُحجم عالم النفس عن الدخول في علاقة متعددة إذا كان متوقعاً بدرجة معقولة أن تضعف العلاقة المتعددة موضوعية عالم النفس، كفاءته، فعاليته في إنجاز وظائفه أو وظائفها كعالم نفس، أو بشكل آخر استغلال الأخطار أو الأذى للشخص الذي توجد معه العلاقة المهنية.

لا تكون العلاقات المتعددة التي من غير المتوقع بصورة مرضية أن تسبب إفساداً أو استغلالاً للخطر أو أذى غير محظورة.

(ب) إذا وجد عالم النفس أنه، وفقاً لعوامل غير مرئية، قد نشأت علاقة مؤذية بصورة محتملة، فهو يتخذ خطوات معقولة لحلها مع أخذه بعين الاعتبار أفضل المصالح للشخص المُتأثر واتفاق مع الدستور الأخلاقي إلى أقصى حد.

(ج) عندما يُطلب من علماء النفس من قبل القانون، السياسة المؤسساتية، أو الظروف غير الاعتيادية الخدمة في أكثر من دور في اجراءات إدارية أو قضائية، في البداية يوضحون توقعات الدور ومدى السرية وفيما بعد كلما تطرأ تغييرات.

4.06 الاستشارات

عند الاستشارة مع الزملاء، (1) لا يكشف علماء النفس عن معلومات سرية قد تؤدي بدرجة معقولة إلى التعرّف على الزبون/ المريض، المشارك في البحث، أو منظمة أو شخص آخر لهم معه علاقة سرية ما لم يحصلوا على موافقة مُسبقة من الشخص أو المنظمة وبغير ذلك لا يمكن تحاشي كشف المعلومات، و(2) إنهم يكشفون عن المعلومات فقط إلى الدرجة الضرورية لتحقيق أغراض الاستشارة. (انظر أيضاً المعيار 4-1، صيانة السرية).

مثال: إن الموضوعات ذات الأهمية للمعيارين 3-5 و4-7 متشابهة وهكذا يتم اعتبارها معاً. أولئك العاملون في تطوير وتكييف الاختبارات ربما يكسبون معرفة بالشركة - الزبون والتي قد تكون ضارة إذا جرى استخدامها خارج ذلك الموقع. على سبيل المثال، شركة - زبون (أ) تعاقدت مع مطور اختبار ليكيّف اختباراً يميل للتأثير بشكل سلبي على مبيعات اختبار مشابه من قبل منافسها (ب). يتم سؤال مطور الاختبار فيما بعد من قبل الشركة (ب) أن يعمل على تحسين اختبارها حتى تحافظ على قيادتها

للسوق. إن المعرفة الأولية لمطور الاختبار والمكتسبة من العمل مع الشركة سوف تساعده أو تساعدها في عملها في الشركة (ب) بدرجة كبيرة. يجري تصميم هذا المعيار بغية التوجه إلى هذه الحادثة وأخرى غيرها والتي تنشأ من أنشطة استشارية ربما تسهم في علاقات متعددة. ربما تكون المجموعات 1-5 (الجدول 1-3) الأكثر تأثراً بتلك المعايير.

3.07 طلبات طرف ثالث للقيام بخدمات

عندما يتفق علماء النفس على أن يقدموا خدمات إلى شخص أو هيئة بناء على طلب طرف ثالث، يحاول علماء النفس أن يوضحوا في بداية الخدمة طبيعة العلاقة مع كل الأشخاص أو المنظمات المعنية. يتضمن هذا التوضيح دور عالم النفس (e.g المعالج، المستشار، المُشخّص، أو شاهد خبرة)، تعريف من يكون الزبون، الاستعمالات المحتملة للخدمات المقدمة أو المعلومات المتاحة، وحقيقة أنه ربما يكون هنالك حدود للسرية.

مثال: ربما تحتاج شركة (i.e، طرف ثالث) معلومات عن طالب وظيفة أو موظف حالي (طرف أول) وهكذا تطلب من عالم النفس ومختص استشاري آخر (طرف 2) أن يديروا التقييم المرجو. ينبغي على (الطرف 2) أن يوضح طبيعة مسؤولياتهم وأولوياتها وأن يعرفوا أنه قبل التقديم الأولي لخدماتهم الاستشارية، المدى الذي يمكن لاكتشافاتهم أن تتواصل مع (الطرف 1).. فوق ذلك، ينبغي (للطرف 2) أن ينقل إلى (الطرف 1) الطريقة التي ربما استخدمها الطرف الثاني في اكتشافات الاختبار وفيما إذا سوف يستلم (الطرف 1) نتائج الاختبار، ويكون قادراً على تحديد الاكتشافات، وأن يتم السماح له بتقديم برهان إضافي. ربما تكون المجموعات 7-10 (جدول 1-3) الأكثر تأثيراً بهذا المعيار.

2.05 تفويض العمل إلى آخرين:

يتخذ علماء النفس الذين يفوضون العمل إلى موظفين، مشرفين، أو مساعدي تعليم أو بحث، أو أولئك الذين يستفيدون من خدمات أشخاص آخرين كالمترجمين،

خطوات مدروسة لـ (1) تجنب تفويض عمل كهذا إلى أشخاص لديهم علاقة متعددة مع أولئك المخدومين التي ربما تقود إلى استغلال أو فقدان الموضوعية: (2) التفويض فقط بتلك المسؤوليات التي من المتوقع أن ينجزها بشكل كامل أشخاص كهؤلاء على أرضية من ثقافتهم، تدريبهم، أو خبرتهم، إما بشكل مستقل أو بمستوى تقديم إشراف؛ و(3) ورؤية أن أشخاصاً كهؤلاء ينجزون مهامهم بشكل كامل.

مثال: غالباً ما يفوض الأشخاص المسؤولين عن تكييف الاختبارات أو تقديم خدمات استشارية تستخدم تلك الاختبارات، غالباً ما يفوضون معاونين بمختلف المهام الفنية. وربما تكون المهام المنجزة من قبل المعاونيين فنية، روتينية، ومستهلكة للوقت ومع ذلك وعلى أية حال لديها تأثير على النوعية المحتملة للخدمات المقدمة. يجري ندب المهام إلى المعاونين فقط عندما يكون المعاونون مؤهلين بشكل جيد لإنجازها وتلقي إشراف مناسبٍ. ربما تكون المجموعات 1-6 (جدول 1-3) الأكثر تأثيراً بهذا المعيار.

6.01 توثيق العمل العلمي والمهني وصيانة السجلات:

يُدع علماء النفس، وإلى درجة ما تكون السجلات في نطاق مراقبتهم، وصيانتهم، ونشرهم، وتخزينهم، وحفظهم، تنظيم السجلات والبيانات المتعلقة بعملهم العلمي والمهني وذلك بغرض (1) تسهيل التزود بالخدمات فيما بعد من قبلهم أو من قبل مهنيين آخرين، (2) السماح باستخراج نسخة مطابقة لتصميم البحث والتحاليل، (3) تحقيق متطلبات مؤسساتية، (4) تأكيد دقة الترتيب والدفعات، و(5) تأكيد الالتزام بالقانون. (انظر أيضاً معيار 4001 صيانة السرية).

6.02 صيانة، نشر، عرض سجلات سرية لعمل علمي ومهني:

(أ) يحافظ علماء النفس على السرية في إنشاء السجلات، تخزينها، الوصول إليها، نقلها وعرضها تحت مراقبتهم، سواء أكانت تلك السجلات مكتوبة أم آلية، أم بأي وسيلة أخرى.

(ب) إذا جرى إدخال المعلومات السرية المتعلقة بُمتلقي الخدمات النفسية في أسس معطيات أو أنظمة سجلات متاحة لأفراد لم يحصلوا على موافقة المُتلقي للوصول إليها، يستخدم علماء النفس الرموز أو تقنيات أخرى لتجنب تضمين المُعرِّفات الشخصية.

(ج) يضع علماء النفس الخطط مسبقاً لتسهيل النقل الملائم، وحماية سرية السجلات والمعطيات في حالة انسحاب علماء النفس من المراكز أو التطبيق.

مثال: تناقش الموضوعات الموصوفة تحت بند 6.01 و6.02 موضوعات مشابهة إلى حد ما، وبذلك يمكن اعتبارها مع بعض. أصبحت أهمية تزويد الأرشيف لعمل الفرد واضحة جداً للمؤلف في أثناء خدمته في هيئة محلفين تراجع ادعاءات عن سوء سلوك أخلاقي لعلماء نفس ومحامين. وغالباً ما يرى عدم القدرة على تقديم برهان مؤيّد لعمل فرد ما كعدم وجود عمل لفردٍ ما. إن الفشل في تقديم أرشيف كاف هو قضية رائدة لمعطيات سوء السلوك الأخلاقي، حتى عندما يبرز احتمال أن تكون تلك التهم غير مبررة.

إن العاملين في تكييف الاختبارات مسؤولون عن توثيق عملهم بشكل تام وذلك لتسهيل فهم الآخرين وتقييم هذا العمل. تشمل هذه الأرشفة الطبيعة الذهنية والنظرية لعملية التكييف، تصميم المشروع، طبيعة المعطيات المكتسبة أثناء هذه العملية، والنظرية لعملية التكييف، تصميم المشروع، طبيعة المعطيات المكتسبة أثناء هذه العملية، المعطيات المستخدمة لتقدير الجدارة والصدق، أسماء، أوراق اعتماد للأشخاص المستخدمين (تشمل المترجمين)، مقدار الوقت المُكرّس لعمل الفرد، وموضوعات مشابهة. يجب أن يجتهد وبشكل مساو أولئك الذين يستخدمون اختبارات مكيّفة في صيانة سجل عملهم.

العناية مطلوبة في الإنشاء، الصيانة، النشر، التخزين، الحفظ، وعرض تلك السجلات. ربما يُطلب من الفرد أن يعرض تلك المعطيات بعد سنوات حول موضوعات لا يمكن تخيلها في الوقت الحاضر. تلوثت سمعة بعض علماء النفس البارزين بعد وفاتهم وتشوهت سمعة العديد من المحترفين الأحياء نظراً لفشلهم في مجال حفظ السجلات هذه.

ينبغي اعتبار الاختبارات المكيّفة مصادر مهنية مهمة وملكية فكرية تضمن الأمان وبالتالي صيانة ملائمة للسجلات (أمن الاختبار، 1999). تشمل هذه السجلات نتائج أولئك المُختبرين باختبارات مكيّفة مع معطيات تعطي برهاناً لمعايير اختبار مكيّف، وصفات سيكولوجية أخرى. ربما تكون المجموعات 1-5 و7-9 (جدول 3-1) الأكثر تأثراً بتلك المعايير.

9.05 بناء الاختبار:

يستخدم علماء النفس الذين يطورون الاختبارات ووسائل تقنية أخرى للتقييم، إجراءات سيكولوجية ملائمة ومعرفة علمية ومهنية سائدة لتصميم الاختبار، والتوحيد القياسي (المعايدة)، والصدق، والتقليل أو استبعاد الانحياز، وتوصيات للاستخدام.

مثال: قبل أن يجري استخدام تقنية الاختبارات المبنية، يجب أن تلبي أقل حد من المعايير (مثال: جمعية البحث التربوي الأمريكية، 1999). ومن المتوقع أيضاً أن تلبي الاختبارات المكيفة الحد الأدنى من تلك المعايير. على أية حال، تملك الاختبارات، المكيّفة خواص سيكولوجية لا تلبي تلك المعايير. على سبيل المثال، ربما تكون المعطيات المعيارية غائبة أو غير ممثلة. وربما تكون التقديرات للجدارة وللصدق المكتسبة على الاختبار المكيّف غائبة أو هزيلة (ضئيلة).

يُحجم علماء النفس شخصياً عن سوء استخدام الوسائل التقنية للتقييم، ولا يُشجعون آخرين على عمل ذلك. ربما يشكل غياب معطيات مطلوبة أو براهين حول

الصفات السيكولوجية للاختبار المُكيّف دعوات غير ملائمة إلى التساؤل حول فائدة وتطبيق مناسب للاختبار وبالتالي لاستخدامه. تجري مقابلة اختبار لهذا المعيار، في قسم منه، بواسطة تطبيق ملائم لطرق تخص اختبارات مكيّفة مطابقة للاختبارات المتقدمة التي أجرها هامبلتون (1994، 2001؛ انظر كذلك الفصل 1، هذا المجلد). تخضع المعرفة المهنية الحالية للتصميم، والمقايسة، والصدق،كما في التعرف على الانحراف وتقليل درجته، تخضع دائماً لتغييرات مستمرة في ضوء نظرية أو بحث جديد. ربما تكون المجموعات 1-5 (جدول 1-3) أكثر تأثراً بهذا المعيار.

9.06 شرح نتائج التقييم:

عند شرح نتائج تقييم ما، بما في ذلك الترجمات الآلية، يأخذ علماء النفس في الحسبان الغرض من التقييم كما هو بالنسبة لعوامل الاختبار المتوعة، قدرات الخضوع للاختبار، وخواص أخرى للشخص الذي يجري تقييمه، كالفروق الثقافية واللغوية، والشخصية، والوضعية، والتي قد تؤثر على أحكام علماء النفس أو تُقلل من دقة تراجمهم. إنهم يشيرون إلى أية حدود ذات مغزى لتفسيراتهم.

مثال: ينطبق القول "الرجل العامل هو جيد كأدواته فقط" على عمل أولئك الذين يقدمون خدمات تقييم عبر استخدام الاختبارات المكيّفة.

المختصون بالتقييم الذين يستخدمون اختباراً مكيّفاً تكون صفاته السيكولوجية مألوفة لهم. إن معرفة صدق الاختبار في ضوء الغرض المقصود الذي يُستخدم الاختبار لأجله، هو وثيق بشكل خاص في تشكيل الأحكام إلى الدرجة التي ربما يكون لدى مستخدم الاختبار ثقة في صنع القرارات المرتكزة على درجات الاختبار النهائية. بالإضافة إلى ذلك، إن استخدام الاختبار مع مجموعات فرعية (مثال: أولئك ذوي عمر معين، جنس، عنصر) حيث تغيب عنها المعطيات السيكولوجية هو غير ملائم. ويكون متوقعاً من أولئك الذي يستخدمون اختبارات مكيّفة أن يعرفوا نوعية أدواتهم، وأن يوصلوا هذه المعلومات بشكل دقيق إلى آخرين عند الطلب، وأن

يشكلوا أحكاماً تخص قابلية تطبيق تلك الأدوات في ضوء برهان البحث. ربما تكون المجموعات 7 و8 (جدول 1-3) الأكثر تأثراً بهذا المعيار.

9.07 التقييم من قبل أشخاص غير مؤهلين

لا يشجع علماء النفس استخدام تقنيات التقويم النفسي من قبل أشخاص غير مؤهلين، ما لم يجر توجيه كهذا لأجل أغراض تدريبية وتحت إشراف ملائم.

أمثلة: تتنوع كفاءات علماء النفس لاستخدام الاختبارات بشكل معتبر في داخل البلدان وفيما بينها معاً (أوكلاند وهيو، 1991). لا يستطيع الفرد أن يتوقع أن يكون كل الأشخاص، الذين يستخدمون الاختبارات، مؤهلين للقيام بذلك. ويتبع بعض الأمثلة.

في داخل الولايات المتحدة، يتطلع العديد من علماء النفس إلى تقوية تطبيقهم المهني، وكذلك مردودهم بتقديم خدمات تقييم بالرغم من حصولهم على إعداد ضعيف في استخدام الاختبار. في داخل أوروبا، تختلف المعايير الأكاديمية والمهنية بشكل معتبر من بلد إلى آخر. وربما يكون لدى علماء النفس في العديد من البلدان الأوربية أيضاً القليل من التدريب في استخدام الاختبار أو انعدام التدريب كلياً. إن إعداد علماء النفس في العديد من بلدان أمريكا الجنوبية يُفضل مظاهر علم النفس النوعية والنظرية ويُعطي إعداداً قليلاً للمظاهر النوعية بما فيها طرائق التقييم.

وغالباً ما ينجم عن تلك الشروط وغيرها كون علماء النفس غير مؤهلين لاستخدام الاختبارات. فوق كل ذلك، إن نقص معرفة ملامح خاصة للاختبار المكيّف الذي يرغبون في استخدامه (مثال: صدقه مع مجموعة معينة) يضعف أيضاً كفاءتهم في اتخاذ أحكام معلومة وحكيمة. ربما تكون المجموعات 7-9 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

9.08 اختبارات قديمة ونتائج اختبار مضى تاريخها

(أ) لا يؤسس علماء النفس تقييمهم أو قراراتهم المعترضة، أو توصياتهم على معطيات أو نتائج اختبار انقضى تاريخها لأجل غرض في الحاضر.

(ب) لا يؤسس علماء النفس قرارات كهذه أو توصيات حول اختبارات ومقاييس مُهملة وغير مفيدة للغرض الحالي.

مثال: يجري غالباً إعادة تشكيل مقاييس القدرات الذهنية (.i.e الذكاء، الاستعدادات الأكاديمية، الإنجاز) كل عشر سنوات لتأكيد استمراريتها، لإضفاء القناعة على أن فروقاً ذات مغزى تطرأ على تلك القدرات خلال هذه المدة من الزمن. مقاييس المزاج، والشخصية، ومفهوم الذات عامة ما يتم إعادة تشكيلها بصورة أقل تكراراً، مُعطية القناعة بأن الفروق ذات المغزى في الزمن المعني في هذه النوعيات لا تحدث بشكل متكرر كسابقه. تخضع الاختبارات المُكيّفة أيضاً إلى مراجعة وذلك للحيلولة دون انقضاء تاريخها وتأكيد سريان مفعولها. ربما تكون المجموعات 1-5 و6-8 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

3.11 خدمات نفسية مُرسلة إلى أو عبر منظمات

(أ) يقدم علماء النفس بشكل مباشر الخدمات إلى مؤسسات أو عبر مؤسسات تُزود الزبائن مسبقاً بالمعلومات وعندما تكون ملائمة يكون أولئك المتأثرين مباشرة بالخدمات حول (1) طبيعة وأهداف الخدمات، (2) المُتلقين المقصودين، (3) من يكن من الأفراد زبائن، (4) العلاقة المستقبلية لعالم النفس مع كل شخص ومنظمة، (5) الاستخدامات المحتملة للخدمات المقدمة والمعلومات المتاحة، (6) من سيكون لديه سهولة الوصول إلى المعلومات و(7) حدود السرية. وحالما يمكن إدراك ذلك، يقدمون المعلومات حول النتائج والقرارات النهاية لخدمات كهذه إلى الأشخاص المناسبين.

(ب) إذا جرى منع علماء النفس بالقانون أو بأدوار منظماتية من تقديم معلومات كهذه إلى أفراد أو جماعات معينة، يبلغون أولئك الأفراد أو الجماعات في بداية الخدمة.

9.10 شرح نتائج التقييم

بغض النظر فيما إذا كان إعطاء الدرجات أو التفسير تم القيام به من قبل علماء النفس، من قبل الموظفين أو مساعدين، أو من قبل خدمات خارجية أخرى أو آلية، يتخذ علماء النفس خطوات معقولة لتأكيد أن إيضاحات النتائج يجري إعطاؤها للفرد أو للمندوب المعين ما لم تعق طبيعة العلاقة تقديم توضيح للنتائج (مثل في بعض الاستشارة المؤسساتية، الغربلة الأمنية أو قبل التوظيف، أو التقييمات الجدلية)، وقد جرى شرح هذه الحقيقة بشكل واضح للشخص الذي جرى تقويمه بصورة مسبقة.

مثال: يتوجه المعياران السابقان إلى الحاجة في كونهما حساسين في نقل المعلومات إلى الزبائن وآخرين غيرهم، وبالتالي التأكيد على أهمية موضوعات اللغة المعنية عند الاختبار ونقل نتائج الاختبار، كما تجري الإشارة إليه لاحقاً، ربما يقلل الفشل في الانتباه إلى موضوعات اللغة المعنية، من تقييم صادق للصفات الهدف.

نادراً ما يجري تقييم قدرات اللغة بشكل مباشر. وتُستخدم اللغة بحذافيرها كعجلة اتصال بغية اختبار صفات شخصية أخرى، ويُستخدم تماماً وظيفة أو أكثر من أربع وظائف للغة (مثال: القراءة، الكتابة، الإصغاء، التكلم)، لتقييم الصفات المعنية (مثال: اهتمامات مهنية، الذكاء، الشخصية). ينبغي أن تقدم ضمانات على أن مهارات الشخص متطورة بدرجة كافية وأن لا تضعف وتقييم الصفات المطلوبة. (كوبيز 1984؛ أوكلاند، برنال، هولى، ناتاليشيون، لنير وريتشارد، 1980).

تكون الحاجة غالباً إلى معرفة صفتين مهمتين للغة عند استخدام اختبارات مُكيّفة: الكفاءة اللغوية والتمكن من اللغة، ضمن أولئك الذين يستخدمون لغتين أو أكثر.

ترجع الكفاءة اللغوية إلى قدرات الشخص لفهم ما يقول الآخرون، للتكلم، للقراءة، وللكتابة. يمكن للشخص أن يُظهر قدرات ضعيفة، عادية أعلى من العادية في وظيفة أو أكثر من وظائف اللغة الأربعة تلك. ينبغي أن لا يجري اختبار أولئك

الذين يشكون ضعفاً في واحدة أو أكثر من الوظائف باستخدام طرق تتظاهر بأن لديها مهارات ملائمة للوظائف الضعيفة. كذلك ينبغي للطرق المُستخدمة لتقديم الخدمات، وتشمل إيضاحات لنتائج الاختبار، أن تكون متطابقة مع المقدرة اللغوية لمتلقي الاختبار. تكون معرفة التمكّن من اللغة مهمة عند تقييم الأشخاص القادرين على استخدام وظيفة أو أكثر من وظائف اللغة الأربعة في لغتين أو أكثر. يعود التمكّن إلى فيما إذا كانت مهارات اللغة لشخص أقل تطوراً، من مثيلاتها، أو أنها أكثر تطوراً في إحدى اللغات منها في لغات أخرى. إحداها تختبر بشكل تام استخدام اللغة الأكثر سيطرة. وربما يحتاج أولئك الذين يظهرون تمكّناً متساوياً في لغتين إلى أن يتم اختبارهم في كلتيهما.

وبشكل عام يوثق الأشخاص البارعون في لغتين أو أكثر المعلومات باللغة التي تم بها الحصول عليها. على سيبل المثال، ربما يكون قد جرى اكتساب صفات شخصية واجتماعية في اللغة الأصلية للشخص بينما ربما يتم اكتساب المهارات الأكاديمية في اللغة الثانية للشخص.

عندما يحدث هذا، ينبغي أن تُقيّم مقاييس للصفات الشخصية والاجتماعية باستخدام اللغة الأصلية للشخص بينما ينبغي لمقاييس الصفات الأكاديمية أن تستفيد من اللغة الثانية.

ينبغي توصيل نتائج الاختبار إلى الزبائن من خلال لغتهم الأكثر تمكّناً وفي ضوء مقدرتهم في استخدام لغتهم المتمكّنة. ربما تكون المجموعات 7-10 (الجدول 1-3) الأكثر تأثراً بتلك المعايير.

9.11 صيانة أمن الاختبار

يشير اصطلاح مواد الاختبار إلى الكتيبات، الأدوات، محاضر المؤتمرات، وأسئلة الاختبار أو الحوافز ولا تشمل معطيات الاختبار كما هي مُعرَّفة بالمعيار 4 . 9، الإفراج عن معطيات الاختبار. يقوم علماء النفس بجهود معتبرة لضمان أمانة وأمن مواد

الاختبار ووسائل تقنية أخرى للتقييم تتماشى مع القانون والالتزام المعقود، وبطريقة تسمح بالالتزام بهذا الدستور الأخلاقي.

مثال: ينبغي اعتبار الاختبارات المكيّفة مصادر مهنية مهمة وملكية فكرية تضمن الأمن. يكون استخدامهم محاطاً بالخطر عند عدم تقييد بيعهم واستخدامهم من قبل مهنيين مُعدين بشكل مناسب، والسماح لأشخاص غير مؤهلين بمراجعة الاختبار، ونسخ الكتيبات والمحاضر، وبطرق أخرى السماح لغير المهنيين بالوصول إلى الاختبار المكيّف. ربما تكون المجموعات 1-5 و7-10 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

5.02 بيانات مقدمة من آخرين:

(أ) يتحمل علماء النفس الذين يستخدمون آخرين لإنشاء أو وضع بيانات عامة تُروج لتطبيقهم المهني، لمنتجاتهم، أو لأنشطتهم، يتحملون المسؤولية المهنية لبيانات كهذه.

(ب) لا يعوض علماء النفس على موظفي الصحافة، والإذاعة، والتلفزيون، أو وسائل الاتصال الإعلامية المهمة مقابل الدعاية في مادة إخبارية.

(ج) يجب لإعلان مدفوعة قيمته يمت إلى أنشطة علماء النفس أن يُعرف أن يمكن التعرّف عليه بوضوح كالتالي:

مثال: إن علماء النفس العاملين في الاختبارات المكيّفة مسؤولون عن الإشراف على الطريقة التي يتم فيها نقل اختبارات كهذه إلى آخرين. يجب أن تكون البيانات التي تقترح اختباراً مكيفاً موازية للمقياس الأساسي، ويعطي درجات نهائية مساوية، وبطرق أخرى اقتراح مساواتها أو صلاحيتها، يجب أن يكون مدعوماً بُمعطيات علمية موثوقة.

إن البرهان البيّن غير كاف. يتخذ علماء النفس خطوات معقولة لتصحيح البيانات المُضلّلة. ربما تكون المجموعات 1-5 و7-9 (الجدول1-3) الأكثر تأثراً بهذا المعيار.

7.01 تصميم برامج للتعليم والتدريب

يتخذ علماء النفس المسؤولين عن التربية وبرامج التدريب خطوات معتبرة لتأكيد أن البرامج قد صُممت لتزود معرفة ملائمة وخبرات مناسبة، ولتقابل متطلبات الترخيص بممارسة العمل، والمصادقة عليه، أو أهدف أخرى لطلبات يُقرّها البرنامج.

مثال: تحتاج البرامج التي تُعدّ أشخاصاً لترجمة الاختبارات إلى مصادر مناسبة. تشمل تلك المصادر لكنها غير مقتصرة على الطلاب والأساتذة مع خلفيات متينة في نظرية القياس التقني والتطبيق (بما فيها تطوير الاختبار)، أساتذة ذوو اطلاع جيد على نظرية وتطبيق تكييفات الاختبار، الأدوات الصلدة والمعرفية المُحتاجة (Hardware and Software)، سوية مع الممارسة (practicum) المدُربة والخبرات العلمية العميقة (Practicum) التي يتعلم فيها الطلاب ترجمة الاختبارات تحت إشراف قدير وقريب. ينبغي أن يعمل مديرو البرنامج على التأكيد على الحصول على المصادر المطلوبة. فوق كل ذلك، نظراً للتغييرات المتوقعة في نظرية ترجمة الاختبار والطرائق، تكون الحاجة إلى جهود لتأكيد الأبعاد التطبيقية والمهنية والأكاديمية لتداول البرنامج. إن المجموعة 6 (الجدول 1-3) هي الأكثر تأثراً بهذه المعايير.

8.02 الموافقة الجوهرية للبحث:

(أ) عندما يُطلب الحصول على موافقة جوهرية، يُعلم علماء النفس المشاركين عن (1) الغرض من البحث، المدة المُتوقعة، والإجراءات؛ (2) حقهم في الامتناع عن الاشتراك والانسحاب من البحث بعد بدء الاشتراك؛ (3)

النتائج المترتبة المتوقعة من الامتناع أو الانسحاب، (5) أية فوائد مستقبلية للبحث؛ (6) حدود السرّية؛ (7) حوافز للاشتراك؛ و(8) من يتم الاتصال به من أجل الأسئلة حول البحث وحقوق المشاركين بالبحث. أنهم يعطون الفرصة للمشاركين المستقبليين لطرح الأسئلة وتلقي الأجوبة.

مثال: حدث تغيير مهم وطفيف عندما طلبت الجمعية الأمريكية لعلم النفس (APA) في دليل النشر: الطبعة الرابعة (1994) من المطبوعات لتشير إلى أولئك الذين يجري علماء النفس البحث معهم كمشاركين، لا كموضوعات بحث. يعطي اصطلاح مشاركين احتراماً أكبر من أجل حقوق أولئك الذين تم اكتساب معطيات الاختبار منهم. عند الاشتراك في جمع المعطيات لاختبارات مُكيّفة، ينبغي أن يتم إعلام المشاركين بأغراض الاختبارات، وأن الاشتراك طوعي، وأنهم سوف لا يتعرضون إلى نتائج عكسية إذا اختاروا عدم المشاركة وينبغي أن يتم إعلامهم أيضاً بالفوائد المحتملة، والمخاطر، وحدود السرية مع اسم وعنوان الأشخاص الذين يتصلون بهم إذا كان لديهم أسئلة أو تعليقات. ربما تكون المجموعات 1-5 و7-11 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

8.11 الغش

لا يقدم علماء النفس أجزاء من عمل أو معطيات شخص آخر كعمل لهم، حتى إذا كان العمل أو المعطيات الأخرى مذكورة عرضياً.

مثال: يحدث الغش عامة في عمل تكييف الاختبار (أوكلاند وهيو، 1991)، خاصة عندما يكون الاختبار مكيّفاً دون موافقة مؤلفيه وناشره. ومن الأرجح أن يكون أولئك الذين يكيّفون اختباراً باستخدام مواد من اختبارات أخرى دون موافقة المؤلفين والناشرين منتهكين للمعايير الأخلاقية. ينبغي أن لا يتم التغاضي عن هذا التطبيق. فوق كل ذلك ربما يخرق هذا التطبيق القوانين في تلك البلدان التي تُؤمن حقوق النشر للملكية الفكرية. كما هو مشار إليه في الجزء الثاني، ينبغي على

المحترفين أن يحصلوا على معلومات إضافية حول التطبيقات المشتبه فيها، وإذا اقتضت الحاجة يتخذون خطوات تضمن عدم استمرارها. ربما تكون المجموعات 4 -8 (الجدول 1-3) الأكثر تأثراً بهذ المعيار.

1.03 نزاعات بين المطالب الأخلاقية والمؤسساتية:

إذا كانت المطالب للمؤسسة التي يعمل معها علماء النفس مُدمجة، أو تلك التي يعملون لأجلها تخالف الدستور الأخلاقي، يوضح علماء النفس طبيعة الخلاف، يعلنون التزامهم بالدستور الأخلاقي، وإلى حد يمكن إدراكه، يحلّون الخلاف بطريقة تسمح بالإلتزام بالدستور الأخلاقي.

مثال: يعمل علماء النفس بصورة متكررة في مواقع لا تتقيد بالدساتير الأخلاقية المتبناة رسمياً. على سبيل المثال، ربما يجري توظيفهم من قبل مؤسسة تعتقد أنه يجب أن يكون لديها اختبار خاص مكيّف لاستخدام مقصود في تاريخ قريب. مهما يكن، ربما لا تسمح المعطيات من الاختبار المكيف بالاستخدام المقصود. على أية حال، ربما يعتقد المدراء ضمن المؤسسة أن اختباراً ما هو أحسن من عدم إجراء اختبار ويقررون استخدام الاختبار المكيّف. يفرض هذا القرار معضلة أخلاقية بالنسبة لعلماء النفس. يتم تشجيعهم للبحث عن حل لهذه المعضلة في وسائل تسمح بالتزامهم بالدستور الأخلاقي. ربما تكون كل المجموعات (الجدول 1-3) متأثرة بهذا المعيار.

1.04 حلّ غير رسمي للانتهاكات الأخلاقية:

عندما يعتقد علماء النفس أنه ربما يوجد انتهاك أخلاقي من قبل عالم نفس آخر، يحاولون حل هذ المسألة بتنبيه ذلك الفرد إذا ظهر أن حلاً غير رسمي ملائم ولا ينتهك التدخل في أية حقوق للسرية.

مثال: ينبغي على المحترف الذي يشك أن تطبيقات ترجمة الاختبار قد تنتهك معايير قانونية أو أخلاقية أن يتشاور مع أولئك العاملين في عمل تكيفات الاختبار للتأكد فيما إذا كانت ثمة انتهاكات تحدث، إذا كان الأمر كذلك، ينبغي على المُحترف

شخصياً، وبواسطة المهنة أن يوقف سير استخدام الاختبار المترجم. وكما يُشار إليه فيما بعد في الحالة (1-5)، ينبغي أن تُؤخذ بعين الاعتبار الجهود لإقرار، إذا كان ذلك متضمناً عقوبة رادع لأولئك الذين ينتهكون المعايير الأخلاقية. ربما تكون المجموعات 1-5 و7-9 (الجدول 1-3) الأكثر تأثراً بهذا المعيار.

1.05 الإبلاغ عن انتهاكات أخلاقية:

إذا سبب انتهاك أخلاقي واضح أذى فعلياً، وربما يؤذي بشدة شخصاً ما أو مؤسسة ولا ينطبق عليه القرار غير الرسمي... أو لم يتم حله بشكل مناسب بتلك الصورة، يتخذ علماء النفس إجراء أبعد يكون ملائماً للحالة. ربما يشمل إجراء كهذا الإحالة إلى الدولة أو إلى لجان متعلقة بالأخلاق المهنية، إلى مجالس مؤهلة في الدولة، أو إلى سلطات دستورية ملائمة. لا يطبق هذا المعيار عندما ينتهك تدخل حقوق السرية أو عندما يجري استبقاء علماء النفس بمراجعة عمل عالم نفس آخر الذي يكون سلوكه المهني موضع تساؤل.

مثال: يتطلب الحكم القضائي لانتهاكات أخلاقية مفترضة من الشخص المتهم أن ينتمي إلى جمعية مهنية التي (أ) لديها دستور للأخلاق مؤسس يبين بالتفصيل المعايير المتعلقة بالانتهاك المفترض و(ب) يقدم نظاماً عملياً للمراجعة والتعزيز، عند وجود سلوك لا أخلاقي ربما يتراوح الناجم المحتمل من عبارات تحذير إلى سحب رخصة الفرد للممارسة وفصل من العضوية في الجمعية.

قد يتم أيضاً اتخاذ إجراء قانوني مدني. ربما تكون كل المجموعات (الجدول 1-3) متأثرة بهذا المعيار.

ملاحظة نهائية

يناقش هذا المجلد موضوعات متنوعة مهمة للاختبارات المكيّفة واستخداماتها. يراجع هذا الفصل ستة مبادئ أخلاقية و25 معياراً من المبادئ الأخلاقية لعلم النفس ودستور السلوك: (2002) لـ APA في ضوء تطبيقات متنوعة والتي ربما تقترن باختبارات مكيّفة واستخداماتها.

تشكل الموضوعات الأخلاقية أحد الأركان الأساسية للتطبيق المهني، ويشمل ذلك المقترن بالاختبارات المكيّفة واستخدامها. وبالتالي يبدو نقاش الموضوعات الأخلاقية المقترنة بالعمل المهم في هذا المجال مضموناً.

وقد تلقي الضوء مناقشات أخرى في المؤتمرات، في المجلات، وفي منتديات عامة أخرى على الحاجة إلى حد أبعد من ترويج معرفة الموضوعات الأخلاقية وتطبيقها الذي ترك أثراً على تطور الاختبار واستخدامه، بما فيها تلك المقترنة بالاختبارات المكيفة. أضف إلى ذلك، ينبغي أن تأخذ المناقشات بعين الاعتبار الحاجة لتطوير دستور أخلاقي يتجاوز الحدود الوطنية ويتناول موضوعات منتشرة وعريضة مهمة لتطبيقات الاختبار. قبل تطوير دستور أخلاقي مرتكز على السلوك يتم ضمان البحث الذي يُعرِّف أنماط سلوك منتشرة ومهمة والتي إما قد تطرح معضلات أخلاقية أو تنتهك بوضوح سلوكاً أخلاقياً.

إن المهنة محظوظة لحصولها على أشكال متنوعة من المنح الدراسية التي تناقش الأخلاق. تشمل تلك المنح التالي: جمعية علم النفس الأمريكية. (2002) الجمعية النفسية البريطانية (1999 a، 1998 b)، لجنة توجيه معايير الاختبار الجمعية النفسية البريطانية (1999)، الجمعية النفسية الكندية (1987)، آيد وآل (1988)، لجنة الاختبار العالمية (2000)، اللجنة المندمجة للتطبيقات الاختبارية (1988)، كندال وآل (1997)، كوين (1997)، كوتشر وكيث سبيغل (1998)، لندسي (1996)، والمجلس الوطني للقياس في التربية (1996). تستطيع الجهود القادمة لاختبار موضوعات الأخلاق أن تستخدم تلك المصادر ومصادر موجودة أخرى كنقطة انطلاق لهذا العمل.

ملاحظات المؤلف

إن توماس أوكلاند هو أستاذ في مؤسسة البحث لجامعة فلوريدا، رئيس سابق للجنة الاختبار العالمية، رئيس سابق لجمعيات مدرسة علم النفس العالمية، ورئيس المؤسسة العالمية لتربية الأطفال.

شكر

يعبر المرء عن التقدير للدكاترة رولاند هامبلتون وفونز فان ووفيجفر لتعليقاتهم البناءة حول المسودة الأولية لهذا المخطوط.

*****

المراجع

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington DC: American Educational Research Association.

American Psychological Association. (1994). *Publication manual of the American Psychological Association: Fourth Edition.* Washington, DC: Author

American Psychological Association. (2002). Ethical principles of psychology and code of conduct. *American psychologist, 57,* 1060–1073.

Anastasi, N., & Urbina, S. (1997). *Psychological Testing: Seventh Edition.* Upper Saddle River, NJ: Prentice Hall.

Bartram, D., & Coyne, I. (1998a). *The ITC/EFPPA survey of testing and test use in countries world-wide.* Technical report for the ITC Council.

Bartram, D., & Coyne, I. (1998b). *The ITC/EFPPA survey of testing and test use in countries within Europe.* Technical report for the EFPPA Task Force.

British Psychological Society. (1998a). *Certificate and register of competence in occupational testing general information pack (Level A).* Leicester, England: Author.

British Psychological Society (1998b). *Code of conduct, ethical principles and guidelines.* Leicester, England: Author.

British Psychological Society. (1999). *Certificate and register of competence in occupational testing general information pack (Level B).* Leicester, England: Author.

British Psychological Society Steering Committee on Test Standards. (1999). *Checklist of competencies in educational testing: Foundation level.* Leicester, England: Author

Byrne, B. (1998). *Structural equation modeling with Lisrel, Prelis, and Simplis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Canadian Psychological Association. (1987). *Guidelines for educational and psychological testing.* Ottawa: Author.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* Clevedon, Avon, England: Multilingual Matters Ltd.

Embretson, S., & Hershberger, S. (1999). *The new rules of measurement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). *Test user qualifications: A data-based approach to promoting good test use. Issues in Scientific Psychology.* Washington, DC: American Psychological Association.

Fitzgerald, C., & Ward, P. (1998). *Computer-based testing: A Global perspective.* An address to the meeting of the International Congress of Applied Psychology, San Francisco, CA.

Haladyna, T. (1999). *Developing and validating multiple-choice test items.* Mahwah, NJ: Lawrence Erlbaum Associates.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.
Hambleton, R. (2001). The next generation of ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164–172.
Herrnstein, D. J., & Murray, C. (1994). *The Bell curve.* New York: The Free Press.
Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children and youth: An international survey. *International Journal of Psychology, 26*(3), 329–344.
International Test Commission (2000). *International guidelines for test-use.* Punta Gorda, Florida: Author
Jensen, A. (1980). *Bias in mental testing.* New York: The Free Press.
Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington DC: American Psychological Association.
Kendall, I., Jenkinson, J., De Lemos, M., & Clancy, D. (1997). *Supplement to guidelines for the use of psychological tests.* Melbourne: Australian Psychological Society.
Koene, C. J. (1997). Tests and professional ethics and values in European psychologists. *European Journal of Psychological Assessment, 13,* 219–228.
Koocher, G. P., & Keith-Spiegel, P. (1998). *Ethics in psychology* (2nd ed.). New York: Oxford University Press.
Lindsay, G. (1996). Psychology as an ethical discipline and profession. *European Psychologist, 1,* 79–88.
Loehlin, J. (1998). *Latent variable models.* Mahwah, NJ: Lawrence Erlbaum Associates.
Mays, V., Rubin, J., Saboruin, M., & Walker, L. (1996). Moving toward a global psychology. *American Psychologist, 51,* 485–487.
McDonald, R. (1999). *Test theory.* Mahwah, NJ: Lawrence Erlbaum Associates.
Mercer, J. (1973). *Labeling the mentally retarded.* Los Angeles: University of California Press.
Muniz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal, and Latin American countries. *European Journal of Psychological Assessment, 15,* 151–157.
National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author.
Oakland, T. (Ed.). (1977). *Psychological and educational assessment of minority children.* Larchmont, NY: Brunner-Mazel.
Oakland, T., Bernal, E., Holley, F., Natalicio, D., Leas, R., & Richard, L. (1980). Assessing students with limited English speaking abilities. *Texas Outlook, 64,* 32–33.
Oakland, T., & Hambleton, R. (Eds.). (1995). *International perspectives on assessment of academic achievement.* Norwell, MA: Kluwer Academic.
Oakland, T., & Hu, S. (1991). Professionals who administer tests with children and youth: An international survey. *Journal of Psychoeducational Assessment, 9*(2), 108–120.

Oakland, T., & Hu, S. (1992). The top ten tests used with children and youth worldwide. *Bulletin of the International Test Commission*, 99–120.

Oakland, T., & Hu, S. (1993). International perspectives on tests used with children and youth. *Journal of School Psychology, 31*, 501–517.

Reynolds, C., & Brown, R. T. (1984). *Perspectives on bias in mental testing*. New York: Plenum.

Rosenzweig, M. (1999). Continuity and change in the development of psychology around the world. *American Psychologist, 54*, 252–259.

Sattler, J. (1988). *Assessment of children: Cognitive applications*. San Diego: Author.

Schumacker, R., & Marcoulides, G. (1998). *Interaction and nonlinear effects in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Standards for educational and psychological testing. (1999). Washington DC: American Educational Research Association.

Test security: Protecting the integrity of tests. [Editorial]. (1999). *American Psychologist, 54*, 1078.

Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology, 23*, 101–117.

4

طرق إحصائية لتحديد العيوب في عملية تكييف الاختبارات

ستيفن ج. سيرغي
جامعة مستشوست/ أمهرست

ليان باتسولا
خدمات الاختبارات التربوية

رونالد ك. هامبلتون
جامعة مستشوست/ أمهرست

عند القيام بأبحاث علمية عبر الثقافات فمن الواجب على الباحثين التدقيق في وسائل التقييم المستخدمة للتأكد من خلوها من أية ترجمات/ تكييفات مثيرة للجدل (ملحوظة مهمة: في مجال التقييم عبر اللغات. يعتبر المصطلح adaptation "تكييف" أجدر بالتفضيل من المصطلح "translation" ترجمة" كونه لا يدل ضمناً على ترجمة حرفية. وتعد عملية تكييف الاختبارات بصورة نموذجية أكثر مرونة حيث تسمح باستعاضات لفظية أكثر تعقيداً بحيث يكون المعنى المقصود مصوناً عبر اللغات. حتى وإن لم تكن الترجمة حرفية برمتها (غيسينغر 1994). في هذا الفصل. يتم استخدام المصطلحين بصورة متبادلة؛ لأن كثيراً من القراء حديثي العهد بهذا المجال يمكن أن يكونوا غير ملمين بالمعنى المراد من المصطلح "adaptation" تكييف.

وبصورة خاصة عندما تقتضي الضرورة مقارنة نتائج اختبار مأخوذة من ثقافات مختلفة. فمن الواجب على الباحثين التدقيق في وسائل التقييم المستخدمة للتأكد من خلوها من انحيازات في المفهوم والطريقة والسؤال (انظر مثلاً فان دي فيفر ولونغ 1997 . فان دي فيفر وتانزر 1997). لأنه في حال وجود مثل هذه الانحيازات ولم يتم تحديدها فإن الاستنتاجات المقارنة عبر الثقافات لن تكون صحيحة؛ لهذا السبب فإن كلاً من مؤلفي "معايير في الاختبار التربوي والنفسي" (جمعية الأبحاث التربوية الأميركية. الجمعية النفسية الأميركية. والمجلس الوطني للقياسات في التربية 1999) و"إرشادات في تكييف الاختبارات التربوية والنفسية" (انظر الفصل الأول من هذا الكتاب) يطالبان الباحثين عبر الثقافات أن يقدموا برهاناً على قابلية المقارنة بين نصوص لغوية مختلفة لتقييم ما عندما يراد أن تكون درجات النصوص المختلفة منه قابلة للمقارنة فيما بينها.

على الرغم من وجود استراتيجيات حكمية (نوعية) وإحصائية لتحديد ومعالجة الانحياز. فإن هذا الفصل يركز على التقنيات الإحصائية في معالجة الانحيازات في المفهوم والطريقة والسؤال التي يمكن ظهورها في عمليات التقييم عبر اللغات. ينقسم هذا الفصل، الذي يتوسع في العديد من الأفكار المطروحة من قبل فان دي فيفر و بورتينغا في الفصل الثاني من هذا الكتاب، إلى ثلاثة أقسام.

يتضمن القسم الأول وصفاً للتقنيات الإحصائية في تقييم تكافؤ المفاهيم. ويعرض القسم الثاني إستراتيجيات لتقييم ومعالجة الانحياز في الطرائق. وفي القسم الثالث، ندرج ونناقش الطرائق التقليدية والحديثة في تحديد الانحياز في الأسئلة. يلقي الجدول 1-4 نظرة عامة على الطرائق والأمثلة المشروحة في هذا الفصل. يتبع هذا الجدول خطة التصنيف المطروحة في فان دي فيفر و تانزر (1997) واللذين قاما بتقسيم مصادر الانحياز الشائعة في عمليات التقييم عبر الثقافات إلى هذه الفئات الثلاثة.

تقنيات إحصائية لتقييم تكافؤ المفاهيم:

يمكن للباحثين عبر الثقافات استخدام التقنيات الإحصائية معاً قبل وبعد الاختبار الميداني لتقدير تكافؤ المفاهيم ووسائلهم التقييمية. وبافتراض عدم توفر اختبارات أو بيانات بدرجة السؤال قبل الاختبار الميداني، فإن الباحثين محصورون ضمن كمية المعلومات الممكن جمعها: لهذا السبب، فإن غالبية الأبحاث حول تكافؤ مفاهيم وسائل التقييم المترجمة قد أجريت على الاختبارات الميدانية والبيانات العملية.

قبل الاختبار الميداني:

في حال عدم توفر بيانات بالإجابات على الأسئلة، فإنه يمكن معاينة تكافؤ مفاهيم نصوص لغوية مختلفة لاختبار معين؛ وذلك بجمع البيانات من خبراء في موضوع البحث يمثلون اللغات والثقافات المختلفة المراد دراستها. وبصورة مماثلة لدراسات صحة المضامين التي تجري في الاختبارات التربوية، فإنه يمكن تصنيف الأسئلة في تقييم معين بالاعتماد على معيار أو أكثر بهدف إلقاء الضوء على المفهوم المقاس. من أحد الأمثلة المبتكرة لاستخدام خبراء في موضوع البحث لتقدير التكافؤ في المفهوم هي الدراسة التي أعدها هيو و تريانديس في العام 1985 . في هذه الدراسة، قامت عينات صغيرة من الحكام المنتمين إلى ثقافات مختلفة بتقدير "التشابه في المعاني لأزواج من الأسئلة المستخدمة في اختبار معين" (صفحة 208). استخدم المؤلفان قياساً متعدد الأبعاد للاختلافات الفردية للكشف عن الخاصيات التي اعتمدها الحكام في تصنيفاتهم للتشابه. في حال كون الخاصيات المستخدمة في تقدير التشابه بين الأسئلة متساوقة لدى جميع الحكام، فهذه دلالة أولية على تكافؤ المفاهيم. وفي حال استخدم الحكام المنتمين إلى ثقافات مختلفة لخاصيات مختلفة، فيمكن الاستفادة من هذه المعلومة في تعديل نسخة أو أكثر من الاختبار.

تبين الدراسة التي أعدها هيو وترنيانديس في العام 1985 إحدى وسائل جمع البيانات عبر الثقافات لتقييم معين قبل البدء بإدارته. وعلى الرغم من توفر أبحاث قليلة في هذا المجال، فإن تصاميم أخرى تستخدم خبراء مضامين من بيئات لغوية

مختلفة، أو خبراء مضامين ثنائيي اللغة أيضاً ممكنة. فعلى سبيل المثال، يمكن الطلب من الخبراء ثنائيي اللغة تقدير التشابه في الصعوبة لأسئلة من اختبار إنجازات. ويمكن لهذا الإجراء تحديد أسئلة يمكن ترميزها لاحقاً لاحتوائها على انحياز إذا ما أجريت دراسات للوظائف التفاضلية للأسئلة (DIF) بعد القيام بإدارة الاختبار.

الجدول 1-4

مصادر الانحياز في تكييفات الاختبارات وبعض المراجع الأساسية

مصدر الانحياز	الوصف	المراجع
الانحياز في المفهوم	المفهوم غير وثيق الصلة في جميع الثقافات (تكافؤ مفاهيمي)، المفهوم غير محدد عملياً بصورة منسجمة عبر الثقافات، قياس المفهوم غير منسجم عبر الثقافات.	
الانحياز في الطريقة	انحيازات في ظروف إدارة الاختبار، عدم الإلمام بأشكال الاختبار في ثقافة أو أكثر، أساليب إجابة تفاضلية (مثلاً: الرغبة الاجتماعية)؛ عدم القدرة على مقارنة العينات (انحياز في الانتقاء)؛ تأثيرات مجري المقابلة (مثلاً: تأثيرات التعميم).	
الانحياز في السؤال	ترجمة خاطئة، وثاقة الصلة التفاضلية للأسئلة عبر الثقافات، عوامل الإزعاج.	

بعد الاختبار الميداني:

بعد الاختبار الميداني، عندما تتوفر بيانات إجابة للممتحنين، فإنه توجد أربعة أساليب إحصائية على الأقل لتقييم تكافؤ المفاهيم عبر وسائل التقييم: التحليل الاستطلاعي للعوامل، التحليل الإثباتي للعوامل، القياس المتعدد الأبعاد، ومقارنة الشبكات المنطقية. نقدم في هذا القسم شرحاً موجزاً لكل أسلوب.

التحليل الاستطلاعي للعوامل:

يعد التحليل الاستطلاعي للعوامل أحد أقدم الطرائق و أكثرها شعبية لتقدير ما إذا كانت النصوص اللغوية المختلفة لاختبار معين تقيس المفهوم نفسه. وفي الحقيقة، إن كلاً من فان دي فيفر و بورتينغا (1991)، و بورتينغا (1991) قد وصفا تحليل العوامل بأنه التقنية الإحصائية الأكثر استخداماً لتقييم ما إذا كان مفهوم معين في ثقافة ما موجوداً بنفس الصورة والتكرار في ثقافة أخرى (مثلاً: بوتشر وغارسيا 1978). يوظف أسلوب التحليل الاستطلاعي للعوامل عنصر تحليل العوامل أو بيانات درجات الاختبار بصورة مستقلة لكل فئة ثقافية. وتعاين مصفوفات تحميل العوامل بعد ذلك نظرياً لمعرفة التساوق عبر الفئات. وبالرغم من كون هذا الأسلوب جذاباً من الناحية الفطرية، فإن مقارنة بنى عوامل مستقلة لا يخلو من صعوبة، ولا توجد أية قواعد متفق عليها بصورة عامة لتحديد متى يمكن اعتبار هذه البنى متكافئة. لهذا السبب، فإن الأساليب الإحصائية خاصة تلك التي بمقدورها استيعاب فئات متعددة معاً، هي أساليب جديرة بالتفضيل. ويعتبر التحليل الإثباتي للعوامل والقياس متعدد الأبعاد ذو الأرجحية الأكثر أسلوبين من هذا النمط.

التحليل الإثباتي للعوامل:

تكون بنية الاختبار في التحليل الإثباتي للعوامل (CFA) مفترضة استنتاجاً وتستخدم بيانات المتحنين لتقدير قابلية النجاح عملياً للبنية المفترضة (انظر مثلاً:

بايرن 1998، 2001، 2003). ويتم دمج هذه البنية المفترضة ضمن نموذج معادلة بنيوية وتجبر أن تكون متساوية عبر جميع الفئات. من إحدى فرضيات تكافؤ المفاهيم النموذجية المختبرة باستخدام التحليل الإثباتي للعوامل (CFA) هي ما إذا كانت مصفوفة تحميل العوامل متكافئة عبر جميع الفئات. وتكون بنية مصفوفة تحميل العوامل عادة "بنية مجموعات مستقلة" (مكدونالد 1985) تنص على أنه: (أ) كل متغير مقاس لا يساوي الصفر في التحميل (Loading) على العامل الذي صمم لقياسه فقط، (ب) العلاقات المتبادلة فيما بين العوامل (أي: الخط القطري السفلي للمصفوفة فاي (Δ) مقدرة بصورة حرة، (ج) الأخطاء المرتبطة بتحميلات العوامل (أي: مصفوفة ثيتادلتا) غير متبادلة العلاقة فيما بينها (مارس 1994).

لقد استخدم الباحثون في ميدان التقييم عبر اللغات التحليل الإثباتي للعوامل (CFA) لتقدير ما إذا كانت بنية العوامل لنص أصلي من تقييم معين متسقة عبر نصوص لاحقة مترجمة إلى لغة أخرى (مثلاً : براون وماركوليديس 1996، سيرسي، باستاري وآلالوف 1998). يعتبر التحليل الإثباتي للعوامل خياراً ملفتاً للانتباه لتقدير تكافؤ المفاهيم عبر وسائل مكيفة نظراً لقدرته على معالجة فئات متعددة معاً، ولتوفر اختبارات إحصائية لملاءمة النموذج، وللتزويد بدلائل وصفية لملاءمة النموذج. عندما يتألف تقييم معين من أسئلة مسجلة نتائجها بصورة ثنائية التفرع، فيمكن أن يكون التحليل الإثباتي للعوامل (CFA) مثيراً للجدل لأن النماذج الأساسية تكون خطية في طبيعتها بينما تكون العلاقات فيما بين الأسئلة ثنائية التفرع غير خطية (مكدونالد 1982) . بالرغم من ذلك، فإنه يمكن التغلب على هذا القصور بنظم هذه الأسئلة مع بعضها ضمن فئات قبل البدء بالتحليل. سيرد مثال على استخدام التحليل الإثباتي للعوامل (CFA) لتقدير تكافؤ مفاهيم نسخ لغوية مختلفة من اختبار معين في قسم لاحق.

القياس متعدد الأبعاد:

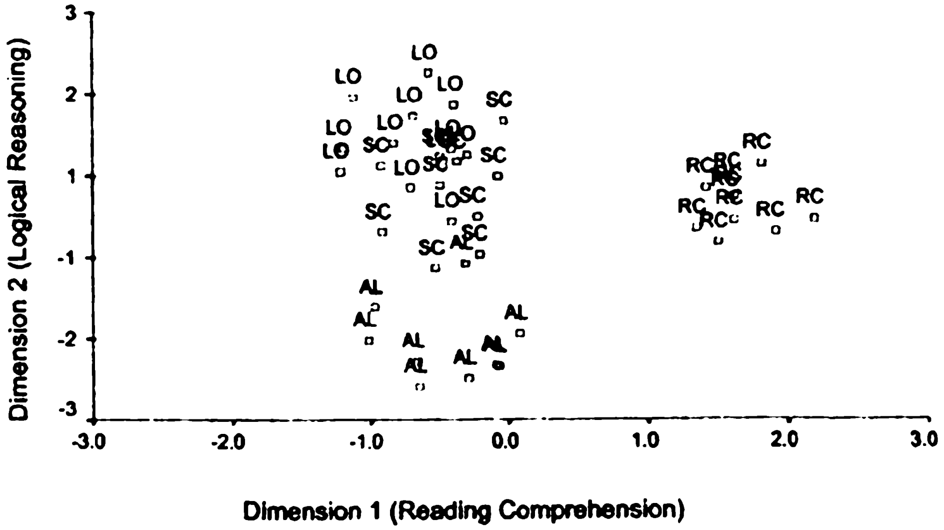
يعتبر القياس متعدد الأبعاد (MDS) أسلوباً آخر مسترعياً للانتباه لتقدير تكافؤ المفاهيم عبر نصوص لغوية مختلفة لاختبار معين.

وبصورة مماثلة للتحليل الاستطلاعي للعوامل، فإن تحليل القياس متعدد الأبعاد لا يتطلب تحديد بنية الاختبار استنتاجياً. بالرغم من ذلك، وبصورة مماثلة للتحليل الإثباتي للعوامل (CFA)، يمكن تحليل البيانات من فئات متعددة معاً . وباستخدام تحليل قياس متعدد الأبعاد للاختلافات الفردية، مثل نموذج (INDSCAL كارول و تشانغ 1970)، يمكن ملاءمة بنية مشتركة مع جميع الفئات معاً، ومن ثم يمكن تقييم الاختلافات البنيوية عبر الفئات بالنظر إلى "أوزان (العينات) الفئات"، والتي تستخدم لتعديل البنية المشتركة لكي تتلاءم بالصورة الأمثل مع بيانات كل فئة. يوفر القياس متعدد الأبعاد وسيلة لكشف الأبعاد التي ترتكز عليها بيانات إجابات الممتحنين، ولتقدير ما إذا كانت هذه الأبعاد متسقة عبر جميع الفئات (أو نصوص الاختبار) المراد دراستها. وهناك ميزة أخرى مثيرة للانتباه للقياس متعدد الأبعاد هي أنه لا يحتاج نموذجاً خطياً لاستنتاج البنية التي ترتكز عليها البيانات.

مثال على التحليل الإثباتي للعوامل (CFA) وتحليل القياس متعدد الأبعاد (MDS) لتكافؤ المفاهيم.

استخدم سيرسي، باستاري، آلالوف (1998) كلاً من التحليل الإثباتي للعوامل (CFA) والقياس متعدد الأبعاد (MDS) لتقدير تكافؤ مفاهيم أسئلة من قسم المحاكمات اللفظية لاختبار القبول بالاعتماد على قياس الذكاء (PET)، وهو عبارة عن اختبار تستخدمه الكليات والجامعات في إسرائيل لاتخاذ قرارات لقبول الطلاب (بيلر 1994 : انظر أيضاً الفصل الثاني عشر من هذا الكتاب). يظهر الرسم التوضيحي 1.4 تمثيلاً ثنائي الأبعاد للأسئلة مستمداً من القياس متعدد الأبعاد (MDS).

تميل هذه الأسئلة إلى الانتظام مع بعضها في فئات في فضاء القياس متعدد الأبعاد (MDS) وفقاً لمواصفات المضامين (التشابهات، المنطق، الفهم عند القراءة، إكمال الجمل). يعد الرسم التوضيحي 4.2 أكثر أهمية، حيث يظهر أوزان الفئات على هذين البعدين ونفسيهما.

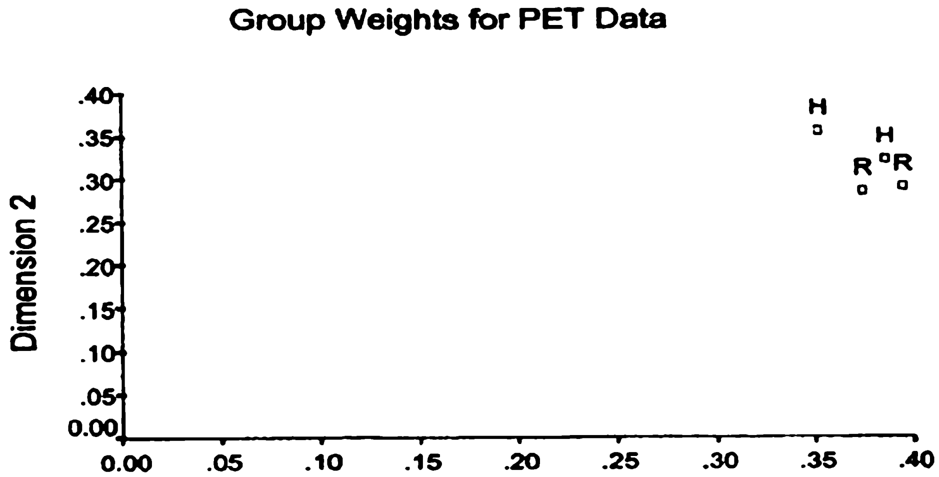


أوزان الفئات على بعدين

لقد تم في هذا التحليل استخدام بيانات فئتين من الممتحنين الذين خضعوا لاختبار النص العبري من الاختبار وفئتين أخريين ممن خضعوا لاختبار النسخة الروسية منه (كانت أحجام العينات حوالي 1300). وكما يتضح من الرسم التوضيحي 4.2، فإن أوزان الفئات كانت متشابهة كثيراً فيما بينها، الأمر الذي يوحي بتشابه البنى (تكافؤ المفاهيم عبر الفئات).

لقد قام سيرسي، باستاري، آالوف (1998) أيضاً باستخدام التحليل الإثباتي للعوامل (CFA) لتقدير تكافؤ مفاهيم هذين النصين المختلفين من هذا الاختبار. وعملاً بمواصفات المضامين، فقد قاموا بملاءمة نموذج رباعي العوامل مع بيانات كلا الفئتين. لقد تم ملاءمة أربعة نماذج مختلفة من التحليل الإثباتي للعوامل (CFA) مع البيانات. قام النموذج الأول بإجبار العوامل الأساسية الأربعة على أن

تكون مشتركة عبر الفئات العبرية والروسية، وقام النموذج الثاني بإجبار مصفوفة تحميل العوامل على أن تكون ذاتها عبر الفئات وقام النموذج الثالث بإجبار الأخطاء المرتبطة بتحميلات العوامل هذه أن تكون ذاتها، أما النموذج الرابع فيحدد أن تكون العلاقات المتبادلة ضمن العوامل متكافئة. تتلخص نتائج تحليلهم في الجدول 2-4. في النماذج الأربعة جميعها، كانت جودة دلائل الملاءمة مرتفعة (96 . أو أكثر) بينما كانت بواقي جذور متوسطات المربعات منخفضة (076 . أو أقل). على الرغم من كون هذه النتائج متسقة مع تحليلات القياس متعدد الأبعاد، فقد تبين لدى استخدام بيانات من تقويم آخر أن الأمور لا تجري دوماً على المنوال نفسه. لذلك فقد أوصوا باستخدام كل من القياس متعدد الأبعاد (MDS) والتحليل الإثباتي للعوامل (CFA) معاً لتقدير تكافؤ المفاهيم لنصوص لغوية مختلفة من اختبار معين.



أوزان المجموعات حسب معطيات PET

مقارنة الشبكات المنطقية:

يعد تكافؤ المفاهيم مصطلحاً عاماً جداً يقول بأن المفهوم النفسي ذاته قابل للقياس عبر جميع الفئات المدروسة وبدقة متساوية في جميع الفئات. يمكن

لأساليب التحليل الاستطلاعي للعوامل، التحليل الإثباتي للعوامل، والقياس متعدد الأبعاد توفير برهان مهم على انسجام بنية الاختبار عبر نصوص لغوية مختلفة لتقييم معين. بالرغم من ذلك فإن البنية المتكافئة لا تقتضي بالضرورة مفاهيم متكافئة. فالتكافؤ البنيوي شرط أساسي ولكنه غير كاف لتكافؤ المفاهيم. لهذا السبب، فقد ارتأى الكثير من الباحثين تجاوز حدود دراسات التكافؤ البنيوي عند تقدير المفاهيم المقاسة عبر نصوص لغوية مختلفة لتقييم معين (مثلاً: فان دي فيفر و تانزر 1997). ويقترح هؤلاء الباحثون اعتماد أسلوب أكثر شمولية يوظف تحري العلاقات فيما بين درجات الاختبار ومتغيرات أخرى مفترضة العلاقة بالمفهوم المقاس.

في المقالة نفسها التي طرح من خلالها المصطلح "صحة المفهوم"، قام كرونباخ وميهل (1955) بطرح المفهوم "الشبكة المنطقية" أيضاً، لقد قاما باستخدام هذا المصطلح لإبراز الحقيقة القائلة بأنه لا يمكن إثبات صدق أي اختبار باستخدام معيار وحيد. بالأحرى، فقد برهنا على أن درجات الاختبار يجب أن تقدر ضمن نظام متعدد المتغيرات يأخذ بعين الاعتبار جميع مظاهر المفهوم المقاس. فيما يتعلق بالتقييم عبر الثقافات، فإن قابلية مقارنة العلاقات المتبادلة بين درجات الاختبار مع متغيرات أخرى يجب أن تكون منسجمة عبر الثقافات، بالإضافة إلى كونها متسقة عبر نصوص لغوية مختلفة بتقييم معين، حتى يكون تكافؤ المفاهيم متيناً. وبالتالي، فإن مقارنة الشبكات المنطقية عبر نصوص اختبارية هو تقييم صارم نظرياً لتكافؤ المفاهيم .

إن مقارنة العلاقات فيما بين درجات الاختبار والمعايير الخارجية المتعددة مهمة يصعب القيام بها في لغة واحدة، تزداد تعقيداً بوجود فئات ثقافية ونصوص اختبارية متعددة. إن تحديد وقياس المتغيرات الخارجية الصحيحة هما مجرد مشكلتين مهمتين علينا التغلب عليهما. لهذا السبب، فإنه من غير المفاجئ ندرة الدراسات الشاملة التي تقارن الشبكات المنطقية عبر الثقافات. بالرغم من ذلك،

فإنه يتوجب على الباحثين عبر الثقافات التدقيق في جدارة كل نص ثقافي لوسائلهم التقييمية والبحث معاً عن براهين صدق متقاربة ومميزة في كل مجموعة ثقافية.

خلاصة نتائج التحليل الإثباتي للعوامل

Model	*GFI*[a]	*RMSR*[b]
Four-factor model common for all groups	.97	.057
Equivalent factor loadings for all groups	.96	.060
Equivalent errors of factor loadings for all groups	.96	.066
Equivalent correlations among factors	.96	.076

[a]GFI = goodness of fit index.
[b]RMSR = root mean square residual

خلاصة:

لا يزال التحليل الاستطلاعي للعوامل أسلوباً شائعاً لتقييم تكافؤ المفاهيم عبر الثقافات. بالرغم من ذلك، يدرك مختصو الاختبارات الحاليون منافع كل من التحليل الإثباتي للعوامل (CFA) والقياس متعدد الأبعاد (MDS) لهذا الغرض. إن استخدام القياس متعدد الأبعاد (MDS) لتقويم تكافؤ المفاهيم عبر الثقافات آخذ بالرواج كونه ممكن الاستخدام قبل وبعد الاختبار الميداني، لا يكون أية افتراضات حول العلاقة فيما بين أسئلة الاختبار، لا يتطلب أن تكون البنية محددة استنتاجاً، وكونه يسمح بتقدير بنية الأبعاد لعدد من الاختبارات في وقت واحد. يعد التحليل الإثباتي للعوامل (CFA) ملفتاً للنظر كونه يمكن استخدامه لإثبات صحة بنية مفترضة معينة و لأنه يوفر نظاماً للاختبار الإحصائي للفرضيات المتنافسة فيما يتعلق ببنية الاختبارات. يوفر كل من القياس متعدد الأبعاد (MDS) والتحليل الإثباتي للعوامل (CFA) معلومات مهمة فيما يتعلق بتساوق بنية الاختبارات. عبر فئات ثقافية مختلفة ونصوص لغوية مختلفة لاختبار معين. بالرغم من ذلك، فإنه

يجب دراسة علاقات درجات الاختبارات بالمتغيرات الأخرى في جميع المجموعات الثقافية المراد دراستها كي نتمكن من تقييم تكافؤ المفاهيم عبر الثقافات على الوجه الأكمل.

استراتيجيات إحصائية لمعالجة وتقييم الانحياز في الطرائق:

بالإضافة إلى تقييم الانحياز في المفاهيم، فإنه يجب على الباحثين أيضاً تقييم الانحياز في وسائل التقييم عبر الثقافات. يشرح فان دي فيفر وتانزر (1997) أن الانحياز في الطرائق نابع من مصادر موجودة في قسم الطرائق للدراسات التجريبية ووفقاً لفان دي فيفر وتانزر فإن هناك ثلاثة أنواع للانحيازات في الطرائق: الانحياز في العينات، في الوسائل، وفي الإدارة. يشير الانحياز في العينات إلى الاختلافات الأساسية عبر الفئات الثقافية أو اللغوية والتي لا علاقة لها بالمفهوم المقاس (مثلاً: الاختلافات في الحوافز على الأداء الجيد، أو الوضع الاجتماعي الاقتصادي). ويشير الانحياز في الوسائل إلى عدم الاتساق في وظائف وسائل القياس عبر الفئات (مثلاً: الإلمام التفاضلي بأشكال الاختبار). أما الانحياز في الإدارة فيشير إلى المشكلات في الإدارة، كإجراءات إدارية غير قياسية (مثلاً: الخطأ في فهم التعليمات الإمتحانية من قبل مديري الاختبار في إحدى الفئات).

نقدم في هذا القسم شرحاً لبعض الإجراءات المتبعة في تقييم الانحياز في الطرائق.

معالجة الانحياز في العينات:

في حال اعتبار المجموعات الثقافية تختلف فيما بينها عبر متغيرات مهمة لا علاقة لها بالمفهوم المقاس، فيمكن استخدام تصاميم أبحاث شاملة وتحليلات إحصائية للتحكم بهذه المتغيرات "المزعجة". حيث يمكن استخدام تحليل مقدار التغاير، تصاميم قوالب عشوائية، وتقنيات إحصائية أخرى (تحليل التراجع، تبادل العلاقة الجزئي إلخ) لعزل تأثيرات مصادر التغيير غير المرغوب بها فيما بين

المجموعات. بالرغم من ذلك، فإن تحليلات كهذه تتطلب جمع بيانات حول هذه المتغيرات الخارجية والتأكد من أن افتراضات الإجراءات الإحصائية قد تم استيفاؤها (مثلاً: تجانس التراجع).

تقييم الانحياز في الوسائل والإدارة:

توجد على الأقل ثلاث استراتيجيات إحصائية لتقييم ما إذا كان الانحياز في الوسائل و/ أو الإدارة موجوداً بين الثقافات قيد الدراسة: الدراسات أحادية الميزة متعددة الطرائق، استخدام معلومات إضافية، واختبار التغيرات. ففي الدراسة أحادية الميزة متعددة الطرائق (فان دي فيفر وتانزر 1997)، يتم استخدام إجراءات تقييم متعددة لقياس الميزة ذاتها عبر المجموعات. وفي حال كون الاختلافات بين الفئات غير متسقة عبر طرائق التقييم المختلفة، فإن تقييماً أو أكثر يمكن أن يكون منحازاً.

يمكن استخدام معلومات إضافية أيضاً لتقدير الانحياز في الوسائل أو الإدارة. توظف هذه الاستراتيجية تحليل متغير ذي علاقة بالمفهوم قيد الدراسة. وفي حال كون الاختلافات الملحوظة عبر الفئات فيما يتعلق بالمعلومات الإضافية غير المتسقة مع الاختلافات الملحوظة فيما يتعلق بدرجات الاختبار، فإن الانحياز في الوسيلة أو الإدارة يمكن أن يوجد. من أحد الأمثلة على استخدام معلومات إضافية للكشف عن الانحياز في الوسيلة و/ أو الإدارة هو استخدام معلومات حول زمن الإجابات، حيث تتم مقارنة مقدار الزمن الذي يستغرقه الممتحنون من فئات مختلفة للإجابة عن سؤال معين (سيرسي، فوستر، أولسن، وروبن 1997). تجعل التقييمات الحديثة والتي تجري بواسطة الحاسب تلك المقارنات أكثر سهولة من أي وقت مضى. فباستخدام اختبارات إحصائية قياسية أو صياغة نماذج معادلات بنيوية أكثر تعقيداً (بايرون 2001)، فإنه يمكن تحديد ما إذا كانت أزمنة الإجابات مختلفة على نحو مهم عبر الثقافات. وفي حال وجود اختلافات عبر الثقافات في أزمنة الإجابات، فيمكن عندئذ السماح بإطالة الحدود القصوى للزمن المعطى لبعض الفئات أو جميعها.

استراتيجية ثالثة لكشف الانحياز في الوسائل و/أو الإدارة تعتمد على إعادة اختبار المتحنين ضمن كل ثقافة (فان دي فيفر وتانزر 1997). حيث يمكن للاختلافات غير المتوقعة الناتجة عن التغيرات بين الاختبار وإعادته عبر الثقافات أن تعكس الانحياز في الوسائل والإدارة (مثلاً: فورمان، يوشيدا، سوانك، وغارسون 1989، فان دي فيفر، دال وفان زونيفيلد 1986). فعلى سبيل المثال، في حال وجود أرباح أكبر في الثقافة (أ) من الثقافة (ب)، فيمكن أن يكون ذلك إشارة إلى أن الثقافة (أ) لم تكن على قدر الإلمام بشكل الاختبار الذي كانت عليه الثقافة (ب) وبالتالي لم يكن أداؤها بنفس الجودة في الاختبار الأولي وحازت على درجات منخفضة فيه. يجب إجراء مثل هذه الدراسات إذا كان هناك أي شك بوجود إلمام تفاضلي بصيغة الاختبار. وفي حال وجود إلمامات تفاضلية كهذه، فإنه يجب أن تتلقى كل ثقافة تعريفاً كافياً بظروف إدارة الاختبار وأشكال الأسئلة قبل أن تتم مقارنة درجاتها.

خلاصة:

يعد تقييم وجود أي انحياز في الطريقة بين ثقافات مختلفة خطوة غالباً ما يتم إهمالها في الدراسات عبر الثقافات. على الرغم من ذلك، فإنها خطوة مهمة. في حال وجود انحياز في الطريقة ولم تتم معالجته، فإن نتائج الدراسة ستكون مضللة. من ناحية أخرى، في حال إمكانية الكشف عن الانحياز في الطريقة ومعالجته باستخدام تحليلات إحصائية أو من خلال تعريف المتحنين بوضع التقييم، فيمكننا عندئذ الانتقال إلى الخطوة التالية في تقدير قابلية مقارنة وسائل القياس عبر الثقافات؛ أي تقييم تكافؤ الأسئلة.

تقنيات إحصائية لتقييم الانحياز في الأسئلة

قبل البدء بمناقشة التقنيات المتبعة في تقييم الانحيازات في الأسئلة، يجب علينا أولاً التمييز بين ثلاثة مصطلحات مهمة ولكنها مستقلة: تأثير السؤال، الوظيفة التفاضلية للسؤال (DIF)، والانحياز في السؤال. يشير تأثير السؤال إلى اختلاف مهم بين الفئات على سؤال معين. على سبيل المثال، عندما تملك إحدى الفئات نسبة أعلى من المتحنين الذين أجابوا على سؤال معين بصورة صحيحة من فئة أخرى. ويمكن لتأثير السؤال أن يكون ناتجاً عن اختلافات حقيقية في الكفاءة بين الفئات أو نتيجة لانحياز فيه. تسعى تحليلات الوظائف التفاضلية للأسئلة (DIF) إلى معرفة ما إذا كان تأثير سؤال معين ناتج عن الاختلافات الكلية بين الفئات في الكفاءة أو نتيجة للانحراف فيه. للقيام بذلك، تتم المطابقة بين ممتحنين من فئتين مراد دراستهما بالنسبة للكفاءة المراد قياسها. يجب على المتحنين ذوي الكفاءة المتساوية والمنتمين إلى فئات مختلفة الإجابة بنفس الصورة على سؤال الاختبار المعطى. في حال عدم الإجابة بصورة مماثلة، فيمكن القول بأن السؤال يؤدي وظيفة مختلفة عبر الفئات.

تعد تحليلات تأثير السؤال والوظيفة التفاضلية له إحصائية في طبيعتها. بينما تكون تحليلات الانحياز في السؤال، من ناحية أخرى، نوعية بصورة أساسية. ويمكن القول بأن سؤالاً معيناً يعتبر منحازاً ضد فئة معينة عندما يكون أداء المتحنين من تلك الفئة أكثر رداءة في الإجابة على السؤال المتصل بالمتحنين في الفئة المرجع والذين هم من الكفاءة نفسها، ويكون سبب الأداء المتدني لا علاقة له بالمفهوم الذي ينوي الاختبار قياسه. لهذا السبب، يشترط لوجود انحياز في سؤال معين تحديد ميزة خاصة بالسؤال تكون غير منصفة لفئة أو أكثر (مثلاً: مفهوم مألوف بصورة أكثر للممتحنين من إحدى الفئات دون غيرها عندما يكون المفهوم ذاته غير ذي أهمية بالنسبة للمهارة المراد تقييمها). وبالتالي، فإن التقنيات الإحصائية لتحديد

الانحيازات في الأسئلة تفتش عن الأسئلة التي تؤدي وظائف مختلفة عبر المتحنين الذين ينتمون إلى فئات مختلفة ولكنهم من كفاءة متساوية. وحالما يتم تحديد هذه الأسئلة، فإنها تخضع إلى اختبار نوعي لتفسير الاختلافات الملحوظة. وعندما يتضح أن لا علاقة لتفسير هذا الاختلاف بالهدف من الاختبار، فإن السؤال يصنف على أنه "منحاز".

نقدم في هذا القسم شرحاً للعديد من أكثر الطرائق شيوعاً والتي تم استخدامها لكشف الانحياز في أسئلة الاختبارات المسجلة درجاتها بصورة ثنائية. يمكن العثور على دراسات أكثر شمولية لاستخدام طرائق الوظائف التفاضلية للأسئلة (DIF) في تحليل البيانات الثنائية في كاميلي وشيبرد (1994)، كلاوسر ومازر (1998)، هولاند و واينر (1993)، ميلسان و إيفرسون (1993)، بوتينزا و دورانس (1995)، وسيرسي وآلالوف (2003). يظهر الجدول 4-3 قائمة بطرائق الوظائف التفاضلية للأسئلة (DIF) في تحليل بيانات الاختبارات. ويوفر هذا الجدول استشهادات بكل طريقة ويشير إلى أنواع البيانات المناسبة لكل طريقة. يحال القراء إلى بينفيلد و لام (2000) بغية الحصول على مراجعة شاملة لطرائق إجراء دراسات الوظائف التفاضلية للأسئلة (DIF) لبيانات إجابات متعددة التفرع.

هناك تطبيقات عديدة لمنهجية الوظائف التفاضلية للأسئلة (DIF) في معالجة مشكلة تقدير الأسئلة المترجمة/المكيفة (مثلاً: ألالوف، هامبلتون، وسيرسي 1999، أنغوف و كوك 1988، بادجيل، رادجو، وكوارتيتي 1995، سيرسي و بيربيروغلو 2000). الطرائق المختارة للمناقشة في هذا الفصل تمثل الطرائق الأكثر استخداماً في المؤلفات العلمية حول تكييفات الاختبارات. الطرائق المبحوثة هي: الرسم البياني للدلتا، التقييس (التوحد القياسي)، مانتل-هاينزل، والطرائق المبنية على نظرية الإجابة على سؤال .(IRT)

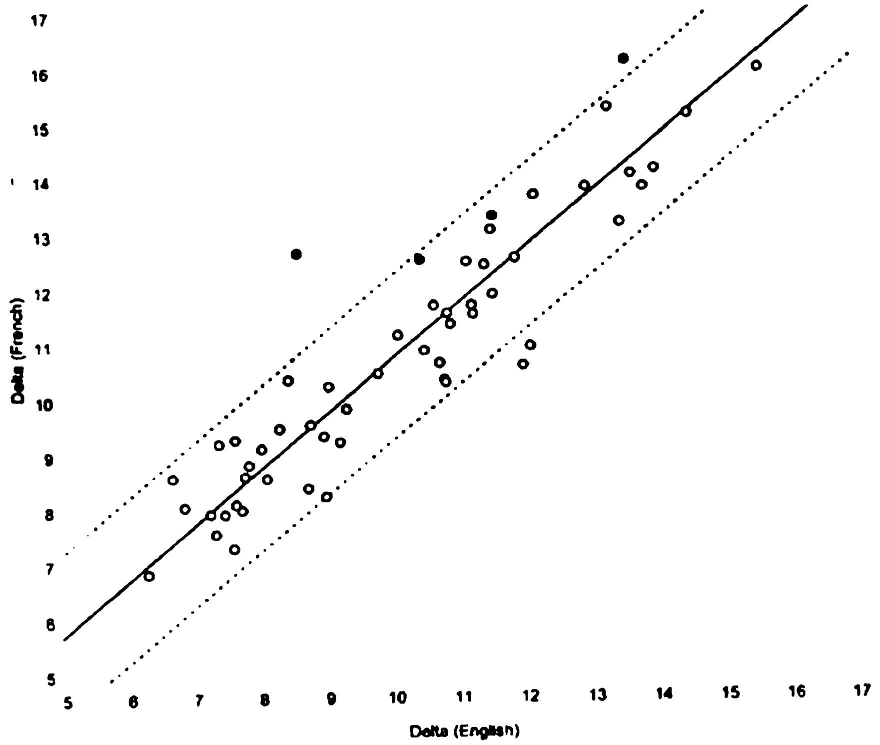
Method	*Sources*	*Appropriate for*	*Applications to Cross-Lingual Assessment*
Delta Plot	Angoff (1972, 1993)	Dichotomous data	Angoff & Modu (1973) Cook (1996) Muniz et al. (2001) Robin, Sireci, & Hambleton (2003)
Standardization	Dorans & Kulick (1986); Dorans & Holland (1993)	Dichotomous data	Sireci, Fitzgerald, & Xing (1998)
Mantel-Haenszel	Holland & Thayer (1988); Dorans & Holland (1993)	Dichotomous data	Allalouf et al. (1999) Budgell et al. (1995) Muniz et al. (2001)
Logistic Regression	Swaminathan & Rogers (1990)	Dichotomous data Polytomous data Multivariate matching	Allalouf et al. (1999) Gierl et al. (1999)
Lord's Chi-Square	Lord (1980)	Dichotomous data	Angoff & Cook (1988)
IRT Area	Raju (1988,1990)	Dichotomous data Polytomous data	Budgell et al. (1995)
IRT Likelihood Ratio	Thissen et al. (1988) Thissen et al. (1993)	Dichotomous data Polytomous data	Sireci & Berberoglu (2000)
SIBTEST	Shealy & Stout (1993)	Dichotomous data	

طريقة الرسم البياني للدلتا:

بالنسبة لأسئلة الاختبارات المسجلة درجاتها بصورة ثنائية (مثلاً: صواب/خطأ)، فإن رسماً بيانياً بسيطاً للتشتت) للإحصائيات الصحيحة للنسبة (قيم النسبة) لكل سؤال غالباً ما يعطي معلومات كافية كاختبار أولي لوظائف الأسئلة عبر اللغات أو الثقافات. (في حالة الإجابات الثنائية على أسئلة نفسية، فإن قيم النسبة تعبر عن نسبة الأشخاص المتفقين مع السؤال). ولإنشاء رسم بياني كهذا، يتم تمثيل قيم النسبة لفئة ثقافية على أحد المحاور، بينما تمثل قيم النسبة لفئة ثقافية أخرى على المحور الآخر. وباستخدام هذين المحورين، فإنه يتم تمثيل كل

سؤال كنقطة في هذا الحيز ثنائي الأبعاد. في حال كون الصعوبات في الأسئلة متسقة عبر الثقافات، فإنها ستقع على طول خط مستقيم بزاوية انحراف قدرها 45 درجة. وبرغم وجود اتساق في صعوبة الأسئلة عبر الفئات الثقافية، فإن بعض التشتت حول الخط المستقيم شيء متوقع نتيجة للأخطاء في أخذ العينات. في حال كون أحد الأسئلة أكثر صعوبة إلى حد كبير (أو أن قلة من الأشخاص يتفقون مع العبارة، في حال كون السؤال مأخوذاً من اختبار نفسي) في ثقافة ما منه في الأخرى، فإنه سيقع بعيداً عن هذا الخط المستقيم، وستتم دراسة هذا السؤال ومثيلاته بصورة إضافية لمعرفة الانحياز المحتمل.

أحد الانتقادات التي توجه إلى طريقة الرسم البياني للتشتت لقيم النسبة بغية تقدير الوظائف التفاضلية للأسئلة (DIF) هو انعدام التحكم بالتأثير. ولأن قيم النسبة هي عبارة عن قيم تابعة للفئة، فمن الصعوبة بمكان إجراء مقارنات واضحة الدلالة لقيم النسبة عبر الفئات. فعلى سبيل المثال، فإن قيم النسبة التي نحصل عليها من فئة ذات أشخاص ذوي كفاءة عالية قد تختلف عن قيم النسبة التي نحصل عليها من فئة ذات أشخاص أقل كفاءة. ويمكن لهذا الاختلاف أن لا تكون له أي علاقة بالانحياز. في حالة كهذه، فإن الاختلافات في قيم النسبة قد لا تكون دالة بالضرورة على عدم تكافؤ الأسئلة عبر الثقافات. وفي الغالب، فإنها ستكون نتيجة للاختلاف في الكفاءة بين الفئات، أو نتيجة للتفاعل بين عدم تكافؤ الأسئلة واختلافات الفئات في الكفاءة. ولمعالجة هذه المشكلة، فقد اقترح آنغوف (1972 و1973) الرسم البياني لـ " قيم الدلتا للأسئلة" لكل فئة عوضاً عن قيم النسبة للأسئلة.



4.3 قيم النسبة للأسئلة

بما أن قيم النسبة للأسئلة عبارة عن قياسات ترتيبية، فمن المألوف اعتبار أن قيم النسبة للأسئلة قد تم الحصول عليها من مجيبين من توزعات مقدرة طبيعية، وإيراد قيم النسبة كانحرافات طبيعية على مقياس ذي متوسط يساوي 13 وانحراف معياري يساوي 4 (يعرف باسم "قيم ETS للدلتا" تيمناً بالمنظمة التي كانت لها الصدارة في استخدامها في مجال تطوير الاختبارات). على هذا النحو، فإن قيمة الدلتا الموافقة لقيمة نسبة 50 . (مثلا) ستكون 13. وإذا كانت قيمة النسبة لسؤال 84. فإن قيمة الدلتا ستكون 9.0 لقيمة نسبة سؤال 16. فإن قيمة الدلتا ستكون 17.0 . وبكل وضوح، فإن قيمة الدلتا للأسئلة الصعبة تكون قيماً مرتفعة، بينما تكون القيم منخفضة في الأسئلة السهلة.لقد جرت العادة على اعتبار الاختلاف في قيمة الدلتا المساوي 1.5 بين فئتين جديراً بالمراجعة الجدية، بعد أخذ أية اختلاف فئوي كلي بعين الاعتبار (هولاند وواينر 1993). عند الرسم البياني التشتتي لقيم الدلتا،

فإن الجزء المحصور من الخط (المساوي طوليا) المار عبر الرسم البياني يمثل الاختلاف الكلي في المقدرة بين الفئات. وتمثل النقاط الواقعة على الرسم البياني ضمن دوائر صغيرة الأسئلة التي لها صعوبة نسبية متساوية تقريباً في كلا الثقافتين. يظهر الرسم التوضيحي 4.3 مثالاً للرسم البياني للدلتا، حيث يبين قيم الدلتا المحسوبة بواسطة مجيبين أجروا اختبار النسخة الفرنسية (المحور العمودي) أو الإنكليزية (المحور الأفقي) لشهادة دولية (من ميونز، هامبلتون، وكروسينغ 2001). وتعبر حقيقة عدم مرور الخط المساوي طولياً من الأصل (نقطة تقاطع محاور الإحداثيات) عن الاختلاف الكلي في الكفاءة بين الفئتين، وهو 77. (لمصلحة الممتحنين باللغة الإنكليزية والذين كان أداؤهم أفضل قليلاً) في هذه الحالة، لقد قمنا برسم نطاق ثقة حول هذا الخط المساوي طولياً وتتم الإشارة إلى الأسئلة الواقعة خارج هذا النطاق على أنها أسئلة ذات وظائف تفاضلية. إن تلك الأسئلة المشار إلى احتوائها وظائف تفاضلية باستخدام إجراءات إحصائية أكثر تعقيداً موضحة أيضاً في الرسم (لمزيد من التفاصيل، انظر ميونز وآخرون 2001) في هذه الحالة، فإن عملية الرسم البياني قد أشارت إلى الأسئلة ذاتها.

تعد طريقة الرسم البياني للدلتا لتقدير الوظائف التفاضلية للأسئلة عبر الثقافات سهلة التطبيق نسبياً ونتائجها سهلة التفسير. بالرغم من ذلك، فقد أظهرنا أن الرسوم البيانية للدلتا تهمل الأسئلة محتملة الانحياز عندما تختلف في قدراتها التمييزية (دورانس وهولاند 1993). لهذه الأسباب مجتمعة، فإن طريقة الرسم البياني للدلتا تقترح كاختبار أولي فقط، أو في تلك الحالات التي تمنع فيها أحجام العينات من إجراء تحليلات إحصائية أكثر تعقيداً. لقد برهن ميونز وآخرون (2001) أن طريقة الرسم البياني للدلتا كانت فعالة في تحديد الأسئلة التي كانت تؤدي وظائف مختلفة جداً عن بعضها عبر الفئات اللغوية، حتى عندما كانت أحجام العينات صغيرة إلى حد 50 شخصاً لكل فئة. لقد تم استخدام الرسوم البيانية للدلتا بصورة فعالية أيضاً مع أحجام عينات كبيرة (آنغوف ومودو 1973، كوك 1996).

دليل التقييس:

لقد تم اقتراح دليل التقييس للكشف عن الوظائف التفاضلية للأسئلة من قبل دورانس وكوليك (1986). ويمكن أن تعرف هذه الطريقة باسم طريقة "قيمة النسبة المشروطة"، حيث تحصى قيم النسبة المستقلة لكل سؤال متوقفة عليه درجة الاختبار الكلية. فعلى سبيل المثال، يمكن مقارنة ممتحنين أجابوا على نصوص لغوية مختلفة من سؤال بالنسبة لدرجة الاختبار الكلية. الفكرة المراد الإشارة إليها هنا هي إمكانية وجود بعض الأسئلة المثيرة للجدل، ولكن إجمالاً، فإن مطابقة ممتحنين من الفئتين اللغويتين طريقة معقولة للعثور على فئات متكافئة من الممتحنين. بعد ذلك، وبالنسبة لممتحنين ذوي درجة اختبار معطاة، يتم حساب نسبة الممتحنين الذين أجابوا على السؤال بصورة صحيحة لكل فئة ومقارنتها. في حال خلو السؤال من أية مشكلات، فإنه يجب على الفئتين ذوات الأداء الكلي المتكافئ أو القريب من المتكافئ أن يكون أداؤهما متساو تقريباً في الإجابة عليه. وتعاد هذه العملية بالنسبة لجميع المستويات الأخرى من درجات الاختبار. من الناحية العملية، تحسب فواصل درجات الاختبار بصورة نموذجية لمطابقة الممتحنين بحيث لا تكون أحجام العينات لكل فاصل درجة اختبار صغيرة جداً (أي: مطابقة كثيفة). ولجعل مهمة ترميز الأسئلة ذات الوظائف التفاضلية أكثر سهولة، فقد اقترح دورانس وكوليك دليل التقييس، والذي يمثل المتوسط عبر الدرجات أو الفواصل الزمنية على مقياس درجات الاختبار لقيم النسبة المشروطة للفئتين. بالنسبة للعينات الصغيرة، يتم أحياناً اختيار خمسة أو ستة فواصل زمنية بين درجات الاختبار. يحسب هذا الدليل (STD-P) بالعلاقة

$$STD-P=\frac{\sum_{m} w_m\,(E_{fm}-E_{rm})}{\sum_{m} w_m}$$

حيث تعبر Wm عن التكرار النسبي للفئة الهدف عند مستوى الدرجة) m أو نسبة الفئة المرجع والفئة الهدف عند مستوى الدرجة، الخيار للباحث). وتكونErm

و Efm نسبة الممتحنين عند مستوى الدرجة m والذين أجابوا عن السؤال بصورة صحيحة في الفئة المرجع و الفئة الهدف على التوالي. يمكن للفئة المرجع أن تمثل الممتحنين الذين أجابوا على النص الأصلي من سؤال معين، بينما يمكن للفئة الهدف أن تمثل الممتحنين الذين أجابوا على النص المكيف منه. أحياناً أيضاً يمكن اختيار الأوزان لتتوافق مع الفئة الهدف، وأحياناً أخرى، يمكن اختيارها لتمثل نسبة الفئتين المرجع والهدف معاً عند مستوى درجة معينة، ويعتمد اختيار الأوزان على اهتمام الباحث الأساسي.

يتراوح دليل التقييس بين -1.0 و1.0 على الرغم من عدم توفر أي اختبار إحصائي مرتبط بالمفردة الإحصائية، فإنه يمكن حساب حجم التأثير. فعلى سبيل المثال، يشير دليل انحراف معياري بقيمة 0.10 وسطياً إلى أن الممتحنين في الفئة المرجع الذين تتم مطابقتهم مع ممتحنين في الفئة الهدف يتجاوزون أداء الفئة الهدف عند كل مسافة منتظمة للدرجة بقيمة 0.10 على مقياس تصحيح النسبة.

لقد تم استخدام قيمة دليل تقييس تساوي 0.10 كمعيار لترميز الأسئلة لاحتوائها على وظائف تفاضلية (مثلاً: سيرسي، فيتزجيرالد، كروسينغ 1998). باستخدام هذا المعيار، فإنه إذا تم ترميز 10 أسئلة في اختبار معين لاحتوائها وظائفاً تفاضلية، وكانت كلها تصب في مصلحة واحدة من الفئتين فقط، فإن المستوى الإجمالي للوظائف التفاضلية للأسئلة في الاختبار سيكون حوالي 1 نقطة على مقياس درجة الاختبار الأولية الكلية لمصلحة الفئة المرجع. وباستخدام بيانات حقيقية أو زائفة، فقد خلص ميونز وآخرون (2001) إلى أن دليل التقييس كان فعالاً في ترميز النصوص المكيفة من الأسئلة لاحتوائها على وظائف تفاضلية عندما تكون أحجام العينات صغيرة.

طريقة مانتل - هاينزل

تعد طريقة مانتل-هاينزل (MH) لتحديد الوظائف التفاضلية للأسئلة شبيهة بدليل التقييس في أن ممتحنين من فئتين مختلفتين يطابقون بالنسبة للكفاءة المراد

قياسها وأن احتمالية النجاح في السؤال تتم مقارنتها عبر الفئات. وتعد طريقة مانتل- وهاينزل (MH) توسعاً في اختبار كاي تربيع للاستقلال (مانتل وهاينزل 1959) إلى الوضع الذي يكون فيه ثلاثة مستويات للتطبق. في محيط الوظائف التفاضلية للأسئلة، تكون هذه المستويات هي: فئة الممتحنين (مثلاً: فئتين لغويتين/ثقافيتين)، فاصل متغير المطابقة (الدرجات التي تتم بالاعتماد عليها مطابقة ممتحنين في فئات مختلفة)، والإجابة عن السؤال (صحيحة أو غير صحيحة). ولكل مستوى من متغيرات المطابقة (بصورة نموذجية، درجة الاختبار الكلية)، يتم تنظيم جدول بأبعاد اثنين في اثنين يصنف فئات الممتحنين بحسب الأداء في الأسئلة. من إحدى المميزات الملفتة للنظر في طريقة مانتل- هاينزل توفر اختباراً إحصائياً للوظائف التفاضلية للأسئلة. بالإضافة إلى توفير اختبار للأهمية الإحصائية، فإنه يمكن أيضاً حساب حجم التأثير وتوجد قياسات تقريبية لتصنيف أحجام التأثير هذه إلى وظائف تفاضلية صغيرة، متوسطة، وكبيرة للأسئلة (دورانس وهولاند 1993). يمكن العثور على تفاصيل حول حساب وتفسير إحصائيات مانتل- هاينزل في هولاند وباير (1988) أو دورانس وهولاند (1993).

تعد طرائق الرسم البياني للدلتا، التقييس، ومانتل-هاينزل شائعة؛ لأنها تحتاج أحجام عينات بسيطة فقط ولا تتطلب برامجيات إحصائية متخصصة لإجراء التحليل. إضافة إلى ذلك، فقد أظهرنا أ ن طريقة مانتل-هاينزل فعالة على الأخص في الكشف عن الوظائف التفاضلية للأسئلة. لهذا السبب، غالباً ما يتم استخدامها كمعيار المقارنة في الدراسات التي تقارن طرائق الكشف عن الوظائف التفاضلية للأسئلة. أحد عيوب هذه الطرائق أنها غير فعالة في تحديد الوظائف التفاضلية "غير المنتظمة" للأسئلة. تصور الوظائف التفاضلية غير المنتظمة للأسئلة الوضع الذي تتغير فيه احتمالية النجاح في سؤال معين عبر الفئات عند نقاط مختلفة على طول مسلسلة الكفاءة. إن الطرائق المبنية على نظرية الإجابة عن سؤال والتراجع النسبي لا تحتوي مواطن الضعف هذه. عيب ثان في هذه الطرائق هو أنها وغيرها

من الطرائق المطروحة في هذا الفصل مقتصرة على البيانات الثنائية. ولحسن الحظ فإن معظم الطرائق في الوقت الحاضر قد تم تعميمها لتعالج بيانات إجابة متعددة الفروع، ولكن تلك الطرائق لن تتم مناقشتها هنا (أنظر مثلاً بينفيلد ولام 2000).

طرق نظرية الإجابة عن سؤال

هناك طرق عديدة للكشف عن الوظائف التفاضلية للأسئلة (DIF) ذات البيانات الثنائية تعتمد «نظرية الإجابة عن سؤال» (انظر هامبلتون، سواميناثان، وروجرز 1991). وبصورة أساسية، تقدر جميع هذه الطرائق إمكانية استخدام مجموعة مشتركة من المقادير متغيرة القيمة لسؤال لشرح وظيفة سؤال معين في كل فئة لغوية/ ثقافية. وفي حال الحاجة إلى مقادير متغيرة مختلفة لشرح وظيفة السؤال في كل فئة، عندها يتم ترميز السؤال لاحتوائه على وظيفة تفاضلية. إحدى طرائق نظرية الإجابة على سؤال للكشف عن الوظائف التفاضلية للأسئلة هي طريقة كاي تربيع التي استخدمها لورد، والتي تختبر المقادير المتغيرة للقدرات التمييزية للأسئلة ومقادير صعوبتها عبر الفئات (لورد 1980). لقد استخدم آنغوف وكوك (1988) هذه الطريقة لتحديد الأسئلة المشتركة المستخدمة في الموازنة بين اختبار أهلية التعليم (SAT) والنسخة الإسبانية منه.

طريقة أخرى مبنية على نظرية الإجابة عن سؤال لكشف الوظائف التفاضلية للأسئلة هي اختبار رادجو للمنطقة بين منحنيين مميزين لسؤال (رادجو 1988،1990). في هذا التحليل، يحسب المنحنى المميز (ICC) لسؤال معين بصورة مستقلة لكل فئة. بعد ذلك، يتم اختبار المنطقة بين المنحنيين المميزين (ICCS) لمعرفة الأهمية الإحصائية. بالنسبة للبيانات المسجلة بصورة ثنائية التفرع، فقد أدخلت هذه الطريقة تحسينات على طريقة كاي تربيع التي استخدمها لورد في أن الاختلافات في الأداء في الأسئلة نتيجة للحدس (أي: المقدار المتغير C) يمكن

تقديرها أيضاً. على الرغم من كون هذه الطريقة قد تم استخدامها غالباً مع فقرات مسجلة بصورة ثنائية التفرع، فإنه يمكن توسيعها إلى الحالة متعددة الفروع.

قام بادجل وآخرون (1995) بمقارنة نتائج الكشف عن الوظائف التفاضلية للأسئلة لطريقة كاي تربيع التي استخدمها لورد، مناطق رادجو المعلمة وغير المعلمة، وإجراءات مانتل-هاينزل عبر الاختبارات العددية والمنطقية التي تم تطويرها باستخدام اللغة الإنكليزية أولاً ثم تكييفها إلى اللغة الفرنسية بعد ذلك. لقد وجدوا درجة كبيرة من الاتساق عبر هذه الطرائق في تحديد الأسئلة ذات الوظائف التفاضلية المهمة.

طريقة شائعة ثالثة مبنية على نظرية الإجابة عن سؤال للكشف عن الوظائف التفاضلية للأسئلة هي طريقة نسبة الاحتمالات (ثيسن، شتاينبيرغ، وواينر 1988، 1993). باستخدام هذه الطريقة، تتم ملاءمة نموذجين مبنيين على نظرية الإجابة على سؤال (IRT) مع بيانات إجابات الممتحنين ويتم تقدير الاختلاف بين ملاءمة هذين النموذجين للبيانات لبيان الأهمية الإحصائية. يكون النموذج الأول الذي تمت ملاءمته مع البيانات نموذجاً لا يحتوي أية وظائف تفاضلية للأسئلة "NO-DIF" حيث يتم استخدام ذات المقادير المتغيرة للسؤال في معايرته في كل فئة أو مجموعة. ويكون النموذج الثاني الذي تمت ملاءمته مع البيانات نموذجاً يحتوي وظائف تفاضلية للأسئلة (DIF)، حيث يتم استخدام مقادير متغيرة مستقلة لمعايرة السؤال في كل فئة. أي أن النموذج الذي لا يحتوي أية وظائف تفاضلية للأسئلة (NO-DIF) يعامل الفقرة على أنها متكافئة عبر الفئات، بينما يعامل النموذج الذي يحتوي وظائف تفاضلية للأسئلة (DIF) السؤال على أنه مستقل في كل فئة. وبصورة واضحة، يكون النموذج المحتوي وظائف تفاضلية للأسئلة أقل محدودية لأنه يدخل في اعتباراته مقادير متغيرة أكثر لملاءمتها مع البيانات. ولتحديد ما إذا كانت هذه المقادير المتغيرة الإضافية ضرورية (أي: هل نحتاج مقادير متغيرة مستقلة لمعايرة

السؤال في كل فئة؟)، فإنه تتم المقارنة بين الاحتمالات المرتبطة بكل نموذج (أي: احتمال الحصول على البيانات في حال كون النموذج صحيحاً). ويتم توزيع الاختلاف بين احتمالات كل نموذج ككاي تربيع (في الحقيقة، إن سجل الاحتمالات هو الذي يوزع ككاي تربيع، ويستخدم عملياً قيمة أقل بمرتين من سجل الاحتمالات لمقارنة الملاءمة عبر النماذج)، وتكون درجات الحرية المرتبطة باختبار كاي تربيع هذا بكل بساطة عبارة عن الاختلاف في عدد المقادير المتغيرة التي تم أخذها بعين الاعتبار في كل نموذج.

لقد تم استخدام اختبار نسبة الاحتمالات المعتمد على نظرية الإجابة عن سؤال بصورة واسعة لتقصي الوظائف التفاضلية للأسئلة (DIF) عبر فئات فرعية أجرت الاختبار بلغة واحدة باستخدام نماذج ثنائية ومتعددة الفروع معاً لنظرية الإجابة عن سؤال (ثيسن، شتاينبيرغ، وجيرارد 1986، ثيسن وآخرون 1988، واينر 1995، واينر، سيرسي، وثيسن1991). في الوقت الحالي، توجد بعض التطبيقات لهذه التقنية في معالجة مشكلة الكشف عن العيوب في تكييفات الأسئلة (سيرسي وبيربيروغلو 2000). من محاسن هذه الطريقة قوتها الإحصائية، مرونتها في معالجة بيانات ثنائية ومتعددة الفروع معاً، وقدرتها على تقدير الأسئلة في أكثر من فئتين في وقت واحد. بالرغم من ذلك، فإن لهذه الطريقة عيباً مهماً يتلخص في أن استخدامها يستغرق وقتاً طويلاً جداً. وتجب لكل فقرة ملاءمة نماذج متعددة معتمدة على «نظرية الإجابة عن سؤال» (IRT) مع البيانات. وعندما يتألف التقييم من عدد كبير من الأسئلة ويتم استبعاد النموذج الذي لا يحتوي أية وظائف تفاضلية للأسئلة (NO-DIF)، فإن عزل الأسئلة ذات الوظائف التفاضلية (DIF) المحددة يصبح عملية شاقة (سيرسي، وبيربيروغلو 2000).

خلاصة:

تتنوع الطرائق الإحصائية في استقصاء الأسئلة المثيرة للجدل نتيجة لعيب في التكييف اللغوي/ الثقافي بين طرائق بسيطة تعتمد التحليل النظري، وطرائق معقدة

تعتمد نظرية القياس الحديثة. يعتمد اختيار طريقة بعينها على عدة عوامل من ضمنها أحجام العينات، عدد الأسئلة الداخلة في التقييم، تحديد درجات الأسئلة، وتوفر برامجيات إحصائية. وفي تلك الحالات المتضمنة أحجام عينات صغيرة نسبياً و أسئلة مسجلة بصورة ثنائية التفرع، فإن طريقتي الرسم البياني للدلتا والتقييس (Standardization) هي الطرائق الموصى بها. وبازدياد أحجام العينات، إن طريقتي مانتل-هاينزل (MH) ونظرية الإجابة عن سؤال (IRT) يمكن أن تكونا جديرتين بالتفضيل. وفي كل الحالات يجب التذكير بأنه قبل إجراء أية تحريات حول العيوب على مستوى الأسئلة فإن الانحيازات في المفاهيم والطرائق يجب أن يتم استبعادها. إن جميع طرائق الكشف عن الوظائف التفاضلية للأسئلة تفترض متغير التطابق المستخدم لمطابقة الممتحنين، ولتكن درجة الاختبار الكلية مثلاً، درجة متغير كامنة (أي: درجة معتمدة على نظرية الإجابة على سؤال)، أو متغيراً خارجاً عن نطاق التقييم، وأنه صالح لغرض المطابقة. إن أي انحياز نظامي في هذا المتغير لن يتم الكشف عنه عند مستويات الأسئلة وسيضعف صحة نتائج الوظائف التفاضلية للأسئلة (DIF) ولتلافي هذا الوضع، ننصح باتباع إجراءات تكييف حذرة (مثلاً: إرشادات تكييف الاختبارات التي تصدرها لجنة الاختبارات الدولية، أنظر الفصل الأول من هذا الكتاب، أو هامبلتون وباتسولا 1999) وإجراء اختبارات إحصائية على الانحيازات في المفاهيم والطرائق.

استنتاجات:

يعتبر تقييم ومقارنة الأفراد الذين يعملون في إطار لغات وثقافات مختلفة تحدياً كبيراً. ولقد أظهرت المقالات النقدية في هذا الفرع العلمي العديد من الأخطار التي تهدد الصدق الداخلي للدراسات عبر الثقافات كالانحيازات في المفاهيم والطرائق والأسئلة. لقد قمنا في هذا الفصل بتلخيص وشرح الطرائق الإحصائية التي يمكن للباحثين استخدامها لتقدير أثر هذه الأخطار على صدق

وسائلهم في التقييم عبر الثقافات. ضمن نطاق خبرتنا العلمية، فإن نتائج تحليلات كهذه يمكن أن تستخدم في إدخال تعديلات على التطور اللاحق للوسائل والذي ستتمخض عنه تقييمات أكثر صحة، ومقارنات أكثر شرعية عبر الأفراد المختلفين في اللغة والثقافة.

*****

المراجع

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Detecting the causes of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1972). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 96–116). Baltimore: Johns Hopkins University Press.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Rep. No. 88-2). New York: College Entrance Examination Board.

Angoff, W. H., & Modu, C. C. (1973). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Rep. No. 3). New York: College Entrance Examination Board.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13,* 12–21.

Brown, R., & Marcoulides, G. A. (1996). A cross-cultural comparison of the Brown Locus of Control Scale. *Educational and Psychological Measurement, 56,* 858–863.

Budgell, G., Raju, N., & Quartetti, D. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal, 56*(8), 472–475.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1,* 55–86.

Byrne, B. (2003). Confirmatory factor analysis. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (Vol. 1). Thousand Oaks, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.
Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart–Young decomposition. *Psychometrika, 35,* 283–319.
Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.
Cook, L. L. (1996, August). *Establishing score comparability for tests given in different languages.* Paper presented at the meeting of the American Psychological Association, Toronto, Canada.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.
Dorans, N. J., & Holland. P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.
Dorans, N. J., & Kulick E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23,* 355–368.
Ellis, B. B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment, 11,* 184–193.
Foorman, B., Yoshida, H., Swank, P., & Garson, J. (1989). Japanese and American children's styles of processing figural matrices. *Journal of Cross-Cultural Psychology, 20,* 263–295.
Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.
Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57–68.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.
Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1,* 1–16.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818–825.

Lord, F. M. (1980). ***Applications of item response theory to practical testing problems***. Hillsdale NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. ***Journal of the National Cancer Institute, 22***, 19–48.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. ***Structural Equation Modeling, 1***, 5–34.

McDonald, R. P. (1982). Linear versus non-linear models in item response theory. ***Applied Psychological Measurement, 6***, 379–396.

McDonald, R. P. (1985). ***Factor analysis and related methods***. Hillsdale, NJ: Lawrence Erlbaum Associates.

Millsap, R. J., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. ***Applied Psychological Measurement, 17***, 297–334.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in test translation. ***International Journal of Testing, 1***, 115–135.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. ***Educational Measurement: Issues and Practice, 19***(3), 5–15.

Poortinga, Y. H. (1991). Conceptual implications of item bias. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.), ***Advances in computer-based human assessment*** (pp. 279–290). Dordrecht, Netherlands: Kluwer Academic.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. ***Applied Psychological Measurement, 19***, 23–37.

Raju, N. S. (1988). The area between two item characteristic curves. ***Psychometrika, 53***, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. ***Applied Psychological Measurement, 14***, 197–207.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. ***Psychological Bulletin, 114***, U552–U566.

Robie, C., & Ryan, A. M. (1996). Structural equivalence of a measure of cross-cultural adjustment. ***Educational and Psychological Measurement, 56***, 514–521.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. ***International Journal of Testing, 3***(1), 1–20.

Shealy, R. & Stout, W. (1993). A model-based standardization differences and detects test bias/DTF as well as item bias/DIF. ***Psychometrika, 58***, 159–194.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. ***Language Testing, 20***(2), 148–167.

Sireci, S.G., Bastari, B., & Allalouf, A. (1998, August). ***Evaluating construct equivalence across adapted tests***. Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Sireci, S. G., & Berberoglu, G. (2000). Evaluating translation DIF using bilinguals. *Applied Measurement in Education, 13*(3), 229–248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses* (Laboratory of Psychometric and Evaluative Research Rep. No. 329). Amherst: University of Massachusetts, School of Education.

Sireci, S. G., Foster, D., Olsen, J. B., & Robin, F. (1997, March). *Comparing dual-language versions of international computerized certification exams.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J., Daal, M., & van Zonneveld, R. (1986). The trainability of abstract reasoning: A cross-cultural comparison. *International Journal of Psychology, 21,* 589–615.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

van de Vijver, F. J., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht, Netherlands: Kluwer Academic.

van de Vijver, F. J., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology, 47,* 263–279.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–186.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

5

استخدام ثنائيي اللغة لتقييم التشابه بين صيغ لغوية مختلفة لاختبار ما

ستيفن ج. سيرسي

جامعة ماساشوستس في آمهيرست

لطالما واجه كل من العلماء المعنيين بقياس سرعة ودقة العمليات العقلية والباحثين التربويين والأطباء السريرين مشكلة تقييم الأفراد الذين يستعملون لغات مختلفة. وفي إطار هذه السياقات، فإن حقيقة وجود عالم متعدد اللغات يعيق استخدام وسيلة تقييم واحدة.

لهذا السبب عُدلت الاختبارات إجمالاً للاستخدام بأكثر من لغة واحدة. إلا أن عملية التكييف لسوء الحظ لا تضمن أن الصيغ اللغوية المتعددة لاختبار ما متعادلة. (انظر مثلاً: فان دي فيجفر و لونغ 2000). وهكذا، تكمن المشكلة الأساسية في تقييم التقاطع اللغوي في تجريد تأثيرات الاختبارات من تأثيرات المجموعة عندما تتم المقارنة بين المجموعات والأفراد الذين خضعوا لصيغ لغوية مختلفة لاختبار ما (لمناقشة مفصلة انظر هاملتون ودي جونغ 2003).

تم النقاش طويلاً أنه عندما يترجم اختبار ما أو (يكيف) من لغة إلى أخرى فإنه لا يمكن اعتبار صيغ اللغتين المختلفتين متعادلة. ويُعزى ذلك إلى الزيادة الحديثة في تقييم التقاطع اللغوي مثلا: بيللر، 1994 ، فوستر، أولسن، فورد و سيريسي، 1997، الجمعية العالمية لتقييم الإنجازات التعليمية، 1994، سيريسي،

إكسينغ و فيتز جيرالد، 1999 قد أعيد تأكيد هذه النقطة من قبل اختصاصيي اختبارات معاصرين عديدين (مثلاً: آنغوف وكوك، 1988، جيسينجر، 1994، هامبلتون،1993، 2002، أوليدو، 1981، بريتو، 1992، سيريسي، 1997، فان دي فيفر و تانزر، (1997). ولكن هذا التحفظ يكاد يكون قديماً قدم ممارسة الاختبارات نفسها. أدرك تيرمان (1916) عدم القدرة على مقارنة درجات النسخة الإنكليزية لستانفورد - بينيه مع التقييم الفرنسي الأصلي لبينية. وبشكل مماثل، حذر ليكرت (1932) في مقاله الإبداعي حول مقاييس النسب الإجمالية من استخدام مقاييس الإتجاهات عبر جماعات "ثقافية" مختلفة. على الرغم من ذلك، فإن العالم اليوم يصبح مكاناً أصغر وفعاليات تقييم تقاطع اللغات تتزايد؛ لذا فإن اتباع طرق يعد شيئاً ضرورياً لتقييم مقارنة الاختبارات المستخدمة عبر لغات متنوعة.

إن البحث المعني بالتقاطع الثقافي (بين الثقافات) والتقييم الأدبي للتقاطعات اللغوية يحتوي على عدة أمثلة إبداعية لتصاميم بحثية وطرق إحصائية صالحة لتقييم مقارنة الاختبارات المترجمة أو المكيفة عبر واحدة أو أكثر من زمرة لغوية. مثلاً: آلالوف، هامبلتون و سيريسي، 1999، بادجيل، راجو وكارتيتي، 1995، إليس وكيميل، 1992، هولين، دراسغو، و كوموكار، 1982، سيريسي، فيتز جيرالد، وإكسينغ، (1998).

لقد استعمل كل من فان دي فيفر وبورتينغا (1997)، فان دي فيفر وتانزر (1997) الدراسات السابقة لتطوير تصنيف أشكال التحيز والتعادل المرتبطة بالتقييم اللغوي التقاطعي (بين اللغات). ويشمل التحيز في هذا التصنيف كلاً من: تحيز المفهوم، تحيز المنهج وتحيز البنود. أما التعادل فيتألف من: التعادل البنيوي، تعادل وحدة القياس وتعادل مدى المقياس كاملاً. وتحدد هذه الفئات أنواع الاستنتاجات المقارنة التي يمكن تطبيقها عبر تقييمات للغات مختلفة. وبشكل أساسي، ومن أجل استنتاجات مقارنة عبر الأفراد الذين اختبروا وفق صيغ لغوية مختلفة لاختبار ما فإنه يجب إثبات صدق هذا الاختبار في كل اللغات.

وعلى الرغم من الجهد الكبير الذي تم بذله لتقدير مصادر الانحياز في تقييم التقاطع اللغوي (انظر مثلاً إلى الدراسات الحديثة التي قام بها آلالوف، 2003، آلالوف و آل، 1999، إيريكان 2002) فقد تم بذل أقل من ذلك بكثير لتقدير التعادل القياسي. واستعملت التصاميم البحثية والتحليلات الإحصائية لتطوير مجموع نقاط متسق عبر الزمر اللغوية، إلا أنه لم يتم تقييم مدى مقدرتهم وقصورهم بشكل كامل. إن الهدف من هذا الفصل هو نقد بعض علوم المناهج في هذا الميدان مع التركيز على تلك التصاميم التي توظف الخاضعين للاختبارات المتمكنين من اللغتين (ثنائيي اللغة) والقادرين عن الإجابة على أسئلة الاختبار سواء "الأصلية" أو "المترجمة/ المكيفة"، وقد تم نقاش نقاط القوة والقصور لخيارات هذه التصاميم البحثية المختلفة اعتماداً على الدراسات في هذا الميدان.

مشكلة إنجاز تعادل معياري شامل في تقييم التقاطع اللغوي

إن المشكلة الكبيرة في تقييم التقاطع اللغوي تُوضح بشكل أفضل عبر مثال. لنفرض أننا أردنا مقارنة الإنجاز الرياضي لمجموعة طلبة يتكلمون اللغة الفرنسية مع مجموعة طلبة يتكلمون اللغة الإنكليزية بالمرحلة الدراسية نفسها في كندا. لنفرض أيضاً أنه تم وضع إطار المنهاج التعليمي نفسه لكل الطلاب في هذه الدراسة الافتراضية. يتم إنشاء نسختين من الاختبار الإنجازي. النسخة الأولى هي الاختبار الأصلي الذي صيغ باللغة الفرنسية أما الثانية فهي النسخة المعدلة باللغة الإنكليزية. للوصول للهدف من هذا المثال سنفترض التعادل في التركيب (أنظر إلى غيرل، 2000، كمثال لكيفية تدقيق التعادل في التركيب). بعد تقديم الاختبارات نلاحظ الفرق في الأداء بين الزمرتين الإنكليزية و الفرنسية. هل تمثل هذه الفروقات التي تم رصدها اختلافاً حقيقياً للمجموعة في الإنجاز الرياضي، أم فرق في الصعوبة بين نماذج الاختبار أم أنها تمثل كليهما؟

لفهم هذه المشكلة بشكل أفضل نتظاهر أنه لدينا مقياس إنجازي "حقيقي" تم على أساسه معايرة كلتا المجموعتين، وأن الزمرة الفرنسية "تفوق الزمرة الإنكليزية أداءً بواسطة انحراف قياسي متوسطياً 0.25 (SD) درجة. لتبسيط الأشياء بقدر الإمكان فإن هذا المقياس الحقيقي للعلامات يشمل الصفر ويشمل انحرافاً قياسياً لرقم واحد. ويتألف هذا الاختبار الرياضي المتخيل من خمسة أسئلة فقط. ونعالج هذه المشكلة باستخدام كل من نظرية الاختبار الكلاسيكية ونظرية الإجابة للسؤال (IRT) انظر مثلاً: هامبلتون، سواميناثان و روجرز، 1991). من المميزات المتعلقة باستخدام نظرية الإجابة للسؤال هي أن درجات المجموعة وإحصائيات صعوبة السؤال يمكن التعبير عنها بالارتكاز على مقياس الدرجات نفسه.

السيناريو1: أولاً، نعتبر أن الحالة هي حيث تكون الأسئلة متعادلة عبر اللغتين الفرنسية والإنكليزية وهذا يتحقق حين تتم ترجمة الأسئلة من الفرنسية إلى الإنكليزية وعملية الترجمة/التكييف لم تغير "الجوهر" الأساسي للأسئلة (مثلاً: الأسئلة متعادلة إحصائياً ولغوياً في كلتا اللغتين). بهذه الفرضية، إذا قمنا بحساب إحصائيات بند نظرية الإجابة للسؤال (IRT) والنظرية الكلاسيكية قد تبدو النتائج مماثلة لنتائج جدول رقم 5-1 . في الجدول رقم 5-1 تظهر بوضوح نتيجتان، الأولى بالنظر إلى القيم (P) نسبة الطلاب الذين أجابوا عن الأسئلة بشكل صحيح، ومتوسط علامات الاختبار، نرى أن المجموعة الفرنسية فاقت المجموعة الإنكليزية أداءً. وهذا استنتاج صحيح. لنتذكر أننا افترضنا أن المجموعة الفرنسية فاقت المجموعة الإنكليزية 0.25 (SD) وحدة. ثانياً إن المتغيرات b (IRT) تقديرات صعوبة الأسئلة) هي نفسها في كلا المجموعتين.

جدول 5-2

إحصائيات توضيحية تعتمد على الطريقة الكلاسيكية ونظرية الإجابة عن السؤال (IRT)

حين يبقى تكافؤ السؤال عبر اللغات

(المجموعة الفرنسية تتفوق "فعليا" بـ 25 وحدة على المجموعة الإنكليزية)

قيم p الكلاسيكية وحدات القياس b المستندة على IRT

Item	رقم p الكلاسيكية فرنسي	إنجليزي	متغيرات d فرنسي	إنجليزي
1	.50	.52	1.25	1.25
2	.60	.62	0.00	0.00
3	.55	.57	.63	.63
4	.65	.67	-0.63	-0.63
5	.70	.72	-1.25	-1.25
Mean Score[a]:	3.0	3.1	0.0	0.25

ملاحظة:

استنتاج 1: تفوقت المجموعة الفرنسية على المجموعة الإنكليزية بالأداء.

استنتاج 2: معايير وحدة قياس سؤال حسب نظرية الإجابة عن السؤال ثابتة عبر المجموعتين.

a: متوسط العلامات مرتكز على معيار العلامة الأولي للإحصائيات الكلاسيكية وعلى معيار الحرف اليوناني الثامن الموحد (1.0) لإحصائيات IRT.

توضح هذه النتيجة ميزة ثبات نموذج وحدة قياس السؤال المعروف لمعايرة IRT هامبلتون وآل، (1991)، أي أن معايير وحدة القياس المستندة على نظرية IRT لا تعتمد على النموذج المستعمل لتحديدها. هذه الخاصية تفسر لماذا تستعمل الطرق المعيارية غالباً لتحديد الاختبارات المقدمة إلى مجموعات لغوية مختلفة (مثلاً

آنغـوف وكـوك، 1988، إيليس، 1989، هـولـين ومــاير، 1986، وودكـوك ومــونوز- ساندوفال، 1993).

يوضح السيناريو 1: أنه حين تكون الأسئلة متكافئة عبر اللغات فإن المشكلات المتعلقة بترجمة المجموعات المختلفة تنعدم. لا يوجد اختلافات تتعلق بالاختبار. وفي حـال مـلاحظة مـثل هذه الاخـتـلافـات فـإنه يمكن إسنادها إلـى الفـروقـات بين المجموعتين اللغويتين. لسوء الحظ فإننا لا نعلم من خلال الممارسة إذا كانت الأسئلة متعادلة عبر اللغات أم لا. لنرى كيف يمكن أن تبدو هذه النتائج في حال كانت الأسئلة أكثر صعوبة في واحدة من الصيغتين اللغويتين للاختبار.

منتدى سور الأزبكية

السيناريو 2: هنا نحن نختلق حالة حيث تكون الأسئلة وسطياً أسهل في الإنكليزية بحوالي 0.25 (SD) وحدة. مع بقاء نتيجة المجموعة (المجموعة الفرنسية تتفوق "فعلياً" بـ 0.25 (SD) وحدة على المجموعة الإنكليزية). يمثل هذا السيناريو الحالة حيث يجعل تكييف الاختبار إلى اللغة الإنكليزية الصيغة الإنكليزية للاختبار أسهل من الصيغة الفرنسية. قد تحدث هذه الحالة إذا، على سبيل المثال، أدخل المترجمون دون انتباه مفاتيح للأجوبة الصحيحة أثناء تعديل الأسئلة أو استعملوا لغة أبسط عبر الاختبار. يمثل الجدول 5-2 النتائج المتصورة لهذا السيناريو.

إن أكثر ملاحظة ملفتة للنظر في الجدول 5-2 هي أن أول استنتاج غير صحيح. عُرفت المجموعة الفرنسية كمتفوقة على المجموعة الإنكليزية. إلا أن معايير النسبة للسؤال"P" ومعدلات المجموعة تشير إلى أن المجموعة الإنكليزية هي المتفوقة. هذا الاكتشاف هو نتيجة لحقيقة أن الأسئلة أكثر صعوبة بالفرنسية منها بالإنكليزية. في هذا السيناريو درجة الاختلاف بين متوسط الصعوبات بالأسئلة عبر اللغتين هي أكبر من درجة الاختلاف بين الإنجاز الحقيقي للمجموعتين (0.50 مقابل 0.25 ، بالتتابع). نشأ استنتاجنا الخاطئ لأن النموذجين الكلاسيكي ونظرية (IRT) لا يفسران حقيقة أن الصيغة الإنكليزية للأسئلة أسهل.

الاستنتاج الثاني هو أيضاً غير صحيح. إن الظروف المتصورة لهذا السيناريو تحدد الأسئلة الإنكليزية على أنها أكثر سهولة، إلا أن معايرة وحدة قياس صعوبة (IRT) لوحدة قياس وحدة قياس (b) هي نفسها لكلا المجموعتين اللغويتين. فكيف يمكن ذلك؟

جدول 2-5

إحصائيات توضيحية تعتمد على الطريقة الكلاسيكية ونظرية الإجابة للسؤال(IRT)

حين تكون الأسئلة الفرنسية أكثر صعوبة

(المجموعة الفرنسية تتفوق "فعليا" ب 25. وحدة على المجموعة الإنكليزية)

	Classical p *Values*		*IRT* b *Parameters*	
Item	*English-French*		*English-French*	
1	.50	.48	1.25	1.25
2	.60	.58	0.00	0.00
3	.55	.53	.63	.63
4	.65	.63	−0.63	−0.63
5	.70	.68	−1.25	−1.25
Mean Score[a]:	3.0	2.9	0.0	−0.25

ملاحظة:

استنتاج 1: تفوقت المجموعة الإنكليزية على المجموعة الفرنسية بالأداء.

استنتاج 2: الأسئلة متكافئة عبر اللغتين (مثلاً: لا يوجد اختلافات).

a: متوسط العلامات مرتكز على معيار العلامة الأولي للإحصائيات الكلاسيكية وعلى معيار الرف اليوناني الثامن الموحد (1.0) لإحصائيات IRT.

يتضح أن معايير وحدة قياس صعوبة ونظرية الإجابة للسؤال (IRT) للصيغتين الإنكليزية والفرنسية متعادلة حيث يظهر أنها ترتكز على مقياس عام (مشترك). بينما هي ليست كذلك. فلم يقم أي من الطلبة الإنكليز بالإجابة عن الأسئلة الفرنسية في حين لم يجب أي من الطلبة الفرنسيين بالإجابة عن الأسئلة الإنكليزية. لهذا فإن معايير وحدة القياس للأسئلة الفرنسية تُحسب بالاعتماد على معطيات الطلبة الفرنسيين فقط، بينما تُحسب معايير وحدة القياس للأسئلة الإنكليزية بالاعتماد على معطيات الطلبة الإنكليز فقط. على سبيل المثال فإن الصيغة الإنكليزية للسؤال (1) تمثل سؤالاً أعلى بحوالي 1.25 (SD) درجة من متوسط صعوبات السؤال لكل الأسئلة الإنكليزية. لا توجد طريقة لمعرفة مدى انحراف هذا السؤال عن متوسط صعوبات السؤال الفرنسي. بطريقة مماثلة، تمثل الصيغة الفرنسية للسؤال سؤالاً أعلى بحوالي 1.25 (SD) درجة من متوسط صعوبات السؤال لكل الأسئلة الفرنسية. على الرغم من أنه لكلا مجموعتي الأسئلة قيمة انحراف 1.25 فهي تمثل انحرافات من متوسط معايير الاختلاف (مثلاً المعيار الإنكليزي والمعيار الفرنسي). هذه القيم الانحرافية المحددة لغوياً غير قابلة للمقارنة عبر اللغات. خذ الآن بعين الاعتبار أن 48 ٪ فقط من طلاب اللغة الفرنسية أجابوا عن السؤال (1) بشكل صحيح، في حين أجاب 50 ٪ من طلاب اللغة الإنكليزية على هذا السؤال بشكل صحيح. ولأن وحدات القياس b كانت نفسها بالصيغتين الإنكليزية والفرنسية فإن النتيجة هي أن متوسط الدرجة للمجموعة الفرنسية على مقياس العلامات منخفض بالمقارنة مع المجموعة الإنكليزية. وقد حدث تعديل مماثل للأسئلة الأخرى. تُزودنا نتائج هذا التحليل باستنتاجات مغايرة لما نعتقده صحيحاً. تبدو هذه الأسئلة متعادلة عبر اللغات (بينما هي ليست كذلك). ويبدو أن الطلاب الإنكليز يؤدون بشكل أفضل من الطلاب الفرنسيين (بينما العكس هو الصحيح).

في السيناريو (2): نحدد كلا المجموعتين واختلافات الأسئلة عبر اللغات. إن السبب في أن تحليلاتنا أثمرت عن استنتاجات خاطئة هو أن نموذج المعيار لم يفسر

هذين العاملين. في الواقع حين تكون الفروقات بين الأسئلة والمجموعات غير معروفة فإنه يجب القيام ببعض الافتراضات. علينا أيضاً أن نفترض أن الأسئلة متعادلة عبر اللغات. ثم نقوم بالبحث عن فروقات بين المجموعات، أو نفترض أن المجموعات متعادلة ثم نبحث عن فروقات الأسئلة. يعكس السيناريو (2) الافتراضات التي حصلت حين تمت معايرة الأسئلة مستعملين الخيارات الافتراضية لبرنامج "IRT" النموذجي مثل بيلوك ميسليفي و بوك، (1990) هذا النوع من التحليل سيقوم بمعايرة وحدات القياس b بشكل مترابط (إلى مقياس شائع) بدون تفسير (صحيح) للفروقات بين المجموعات لتصل إلى نتائج كتلك المتمثلة في الجدول (5-2) وهي معقولة تماماً. إن استخدام طريقة قياس تحويلية كما في Stocking و(1983) Lord، التي تعدل وحدات القياس من معايرة ما لتكون على نفس المقياس كأسئلة من معايرة مختلفة لا يمكن أخذها بالاعتبار لأنه لا توجد أسئلة متاحة غير شائعة (مثلاً: غير مترجمة) لإجراء تعديل كهذا.

هل يعد السيناريو (2) واقعياً؟ على الأغلب أنه ليس كذلك. بإعطاء طرق تكييف وترجمة اختبار دقيقة (للتطورات في طرق تكييف الاختبار، انظر إلى هاملتون، المقطع (1)، هذا المجلد: هاملتون وباتسولا، 1999، ومـوليس، كيللي، وهالي، 1996، فإنه من غير المحتمل أن كل الأسئلة في صيغة لغوية واحدة لاختبار ما ستكون أكثر صعوبة من نظيراتها في صيغ لغوية أخرى. إن سيناريو أكثر احتمالاً سيحوي بعض الأسئلة الأكثر صعوبة في الفرنسية وأخرى أكثر صعوبة بالإنكليزية. على أية حال فإن النقطة المهمة هي أن عدم تعادل السؤال و عدم تعادل المجموعة يمكن، وعلى الأغلب أنها كذلك، أن تظهر في الوقت نفسه في تقييم التقاطع اللغوي. حين يتوفر هذان العاملان، فإن مناهج القياس التقليدية غير كافية لاستخراج استنتاجات حول الفروقات بين الاختبارات والمجموعات عبر اللغات. إن الشيء الضروري للقيام بمثل هذه الاستنتاجات هو إما طريقة لتفسير اختلافات المجموعة ضمن المعايرة أو نماذج الدرجات، أو مجموعة أسئلة يمكن اعتبارها متعادلة عبر

اللغتين. كما سنشرح لاحقاً، فإن الطريقة لتحديد أسئلة يمكن اعتبارها متعادلة عبر اللغات هي تقديم الاختبارات إلى ممتحنين ثنائيي اللغة.

استخدام ثنائيي اللغة لتقييم صيغ لغوية مختلفة لاختبار ما:

واحدة من الطرق لمعالجة تعادل صيغتين لغويتين مختلفتين لاختبار ما هي تقديم الصيغتين اللغويتين المنفصلتين إلى مجموعة من الخاضعين للاختبار المتمكنين من كلا اللغتين (ثنائيي اللغة). المنطق الأساسي لهذا الاختبار هو أنه باستخدام مجموعة واحدة من الخاضعين للاختبار "مجموعة لغوية"، فإنه يتم إهمال النتائج، ويمكن تحقيق تعادل معياري كامل. وهكذا فإن رصد الفروقات لأداء السؤال أو الاختبار عبر اللغات يمكن أن يعزى إلى الاختلافات اللغوية بين الاختبارات أو الأسئلة. على الرغم من أنه باستخدام مجموعة واحدة عادة ما تهمل اختلافات المجموعة في أغلب التصميمات البحثية، فإن هناك بعض الخلل في هذا المنطق حين يطبق على مسألة تقييم التقاطع اللغوي. إن المشكلة الأكثر ظهوراً هي الافتراض المطلق أن ثنائيي اللغة متمكنين من كلتا اللغتين بالدرجة نفسها. على سبيل المثال في حال أدت مجموعة من ثنائيي اللغة بشكل مختلف في الصيغ اللغوية " A وB" لسؤال ما، فإن نسب هذا الاختلاف إلى التكييف الخاطئ يفرض أن ثنائيي اللغة سيؤدون بالطريقة نفسها في كلتا الصيغتين اللغويتين للسؤال إذا كان التكييف وافياً. على الرغم من ذلك، فمن المحتمل أن الممتحنين الثنائيي اللغة هم أكثر تمكناً في لغة من الأخرى. لذا، توجد فرضية منافسة معقولة وهي أن ثنائيي اللغة يؤدون بشكل أفضل في الأسئلة المقدمة بلغتهم الأقوى، حتى حين تكون نسختا السؤال متعادلتين بحق.

خلل آخر يتعلق بهذا المنطق هو أنه يصف ثنائيي اللغة كما لو أنهم أحاديو اللغة، مجموعة متجانسة من الخاضعين لاختبار، بينما في الحقيقة، إن مجموعة ثنائيي اللغة الخاضعين لاختبار تتألف على الأغلب من أفراد ذوي خلفيات ومقدرات ومهارات لغوية مختلفة بيكر، 1988، فالدي وفيجورو، (1994). في الولايات المتحدة،

على سبيل المثال، فإن مجموعة من ثنائيي اللغة الإنكليزية - الإسبانية يمكن أن تشمل أناساً لغتهم الأولى الإنكليزية وتعلموا الإنكليزية في المدرسة الثانوية و مهاجرين (من عدة بلدان متنوعة) ناطقين باللغة الإسبانية قد تعلموا حديثاً التكلم باللغة الإنكليزية و مهاجرين من الجيل الثاني تعلموا الإنكليزية كلغة ثانية في المدرسة الابتدائية. لهذا، فالفرضية بأن ثنائيي اللغة يمثلون "نمطاً" واحداً من الخاضعين لاختبار غير منطقية. كما سأناقش لاحقاً، إن الاختلافات اللغوية ضمن مجموعة من ثنائيي اللغة يجب أن تدرج في التصميم البحثي عند استخدام ثنائيي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة. إن المشكلة الأكثر دقة ولكن الجدية، هي أن استخدام ثنائيي اللغة من أجل تقييم الاختبارات هي المقارنة المثيرة للجدل لثنائيي اللغة و أحاديي اللغة (هاملتون وكانجي، 1995) تم تسمية هذه المشكلة سابقاً المشكلة التمثيلية (سيرسي، 1997). في الاختبارات التعليمية، مثلاً، يميل ثنائيو اللغة للاختلاف عن مجموعاتهم الأحاديي اللغة. قد يكون ثنائيو اللغة الذين يجيدون لغتين بكفاءة عالية ممثلين فقط للطلاب الأعلى إنجازاً في المجموعة الأحادية اللغة أيضاً. بشكل معاكس، فإن ثنائيي اللغة الذين يجيدون لغة أو لغتين بشكل هامشي يمثلون فقط الطلاب الأقل إنجازاً في واحدة من المجموعات الأحادية اللغة. على أية حال، فإن تصنيف الكفاءة في نموذج ثنائيي اللغة تميل إلى أن تكون مختلفة جداً عن التصنيفات المقابلة لجماعاتهم أحاديي اللغة.

على الرغم من أن استخدام ثنائيي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة تتطلب البراعة حين تخصص للتصاميم البحثية وحين تستعمل البيانات التحليلية لمعطيات الحالة الفنية، فقد يزودنا ثنائيي اللغة بمعلومات قيّمة فيما يتعلق بتعادل الاختبار ومقارنة الدرجات. تم طرح خيارات التصاميم البحثية في إشعار هذا المقطع لاستخدام ثنائي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة، وتم ملاحظة متغيرات مربكة بحاجة إلى السيطرة عليها. بالإضافة إلى أنه تم التزويد باقتراحات لاستخدام ثنائيي اللغة للقيام بتقييمات أكثر شمولاً للاختبارات المقدمة بعدة لغات.

بدائل عن التصاميم البحثية باستخدام ثنائيي اللغة

تصاميم أحادية المجموعة:

يتطلب التصميم ثنائي اللغة للمجموعة الأحادية تقديم صيغتين لغويتين مختلفتين لاختبار ما لمجموعة واحدة من الخاضعين لاختبار ثنائي اللغة. في هذا التصميم لا يوجد تطابقات ضمن المجموعة ثنائية اللغة المُعَدة لتحديد "أنماط" مختلفة لثنائيي اللغة. إلا أن إدارة الاختبارات اللغوية المختلفة عادة ما تكون متعادلة. حيث إن نصف الممتحنين تقريباً يخضعون للاختبار في اللغة A أولاً بينما يخضع النصف الآخر للاختبار في اللغةB. يشابه هذا الاختبار تصميم المجموعة الأحادية المشروح في اختبار الدراسات السابقة المتعلقة بالتعادل. (مثال: كولن وبرونان، 1995).

من الممكن أن يستخدم التصميم ثنائي اللغة للمجموعة الأحادية لتعديل درجات اختبار بأكملها على صيغ لغوية مختلفة. على سبيل المثال، قام بولدت (1969) بحساب مقارنة العلامات للصيغتين اللغويتين الإنكليزية والإسبانية المتعلق باختبار الأهلية للمدارس الثانوية (SAT) عبر اختبار مجموعة صغيرة (عدد= 14) لثنائيي اللغة الإسبانية - الإنكليزية من طلاب المدرسة الثانوية بواسطة صيغتي الاختبار، استنتج أن طرح 200 درجة من درجة اختبار طلاب اللغة الإسبانية يزودنا بتقدير علامة الطلاب المتوقعة لاختبار اللغة الإنكليزية.

يمكن استخدام هذا التصميم أيضاً لتقييم أداء الأسئلة الفردية عبر اللغتين. على سبيل المثال، إن أداء ثنائيي اللغة في كل سؤال يمكن أن يستخدم لتقييم وظيفية الأسئلة التباينية (DIF) عبر اللغات. في حال أدت الأسئلة بشكل متباين في اللغتين (فيما يتعلق بإحصائيات السؤال كصعوبة السؤال أو التمييز)، فإن الأسئلة تصنف بوصفها تعمل بشكل مختلف عبر اللغات، وهي لا تستعمل لإرساء الاختبارات على مقياس عام. بدلاً من ذلك، فإن الأسئلة التي تظهر إحصائيات متشابهة عبر اللغتين ليست أسئلة "DIF" يمكن أن تستخدم في تصميم موازنة إرساء الاختبار لتثبيت أو ربط الاختبارات إلى مقياس عام. إن كلاً من طريقتي النظرية الكلاسيكية ونظرية (IRT) يمكن أن توازنا المعايير بهذا النمط.

هناك ثلاثة نقاط ضعف أساسية على الأقل في التصميم ثنائي اللغة للمجموعة الأحادية. من الواضح أن التصميم لا يفسر وجود أنماط مختلفة من ثنائيي اللغة من الخاضعين للاختبار. إذا تم إجراء الدراسة باستخدام ثنائيي اللغة الذين تهيمن عليهم اللغة A، فإن النتائج قد لا تعمم للحالة التي يستخدم فيها ثنائيي اللغة الذين تهيمن عليهم اللغة B. نقطة الضعف الثانية هي وجود أثر الممارسة. لأن الممتحنين يجيبون عن كل سؤال في كلا اللغتين، فإن الإحاطة بالسؤال في نموذج الاختبار الأول قد يؤثر على إجابات الممتحنين إلى السؤال المماثل في نموذج الاختبار الثاني. بالرغم من أن عامل الموازنة قد يضبط هذا على المعدل، إلا أن النتائج قد تختلف عما يمكن ملاحظته في حال أجاب الممتحنون على نموذج اختبار واحد. أما نقطة الضعف الثالثة فهي أن الطريقة غير اقتصادية نسبياً. إن اختبار مجموعة واحدة من الخاضعين لاختبار ذي نموذجين يضاعف وقت تقديم الاختبار اللازم لإكمال الدراسة. نقطة ضعف مرتبطة بذلك هي أن الممتحنين قد يفقدون الحافز أو يصبحون أكثر إرهاقاً من أن يخضعوا لنموذج ثان مشابه للأول.

مثال جدير بالذكر للتصميم الثنائي اللغة للمجموعة الأحادية هو الدراسة التي أجريت لربط التقييم الإسباني للتعليم الأساسي (SABE) بنظيراتها في اللغة الإنكليزية واختبار كاليفورنيا الإنجازي (CAT) ، والاختبار الشامل للمهارات الأساسية (CTBS) (1988 ، ماك - غراو هيل /CTB) هناك ميزات عدة لهذه الدراسة مثيرة للإعجاب. أولاً، من أجل الضمان أن الطلاب ثنائيي اللغة ماهرون في كلا اللغتين، تم استخدام تقييمات المدرس واختبارات المهارة اللغوية لتصفية الطلاب الذين لم يكونوا متمكنين في اللغتين الإنكليزية والإسبانية. ثانياً، بدلاً من جعل الطلاب ثنائيي اللغة يخضعون لاختبارين منفصلين، قُدم للطلاب مجموعات أقصر من الأسئلة الإنكليزية و الإسبانية المعتمدة. وقد استُخدم أداء ثنائيي اللغة في هذه الأسئلة المعتمدة لاشتقاق جداول تحويلية لمقارنة أداء الطلاب في الـ SABE مع أداء الطلاب في CAT و CTBS.

على الرغم من أنه تم توظيف التصميمات البحثية المحددة في دراسة الـ SABE التي عالجت بعض النقائص لتصميم المجموعة الأحادية، إلا أنه باستخدام أكثر من مجموعة واحدة من ثنائيي اللغة يمكن تحسين الصدق الداخلي والخارجي لهذا النوع من الدراسة.

تصميمات متعددة - المجموعات

تصميم ثنائي - المجموعة. إن التصميم الأكثر بساطة للمجموعة المتعددة ثنائية اللغة يستخدم بشكل عشوائي مجموعتين متعادلتين ثنائيتي اللغة. يمكن إنشاء هذه المجموعات من خلال تمرير صيغتي اختبار أو تعيين ممتحنين لهذه الصيغ بشكل عشوائي. في هذا التصميم، تخضع كل مجموعة لواحد أو اثنين من صيغ الاختبار، مع إلغاء أي احتمال لأثر الممارسة. بالإضافة إلى هذا، فإن كون المجموعتين متعادلتين بشكل عشوائي، يحتم انعدام أي تأثير على المجموعة. هذا التصميم أيضاً اقتصادي أكثر من تصميم المجموعة الأحادية. يمكن جمع بيانات صيغتي الاختبار خلال مقدار الوقت الذي يستغرقه تقديم صيغة اختبار واحدة.

إنشاء صيغتي اختبار. إن نمط الاختبار لكل مجموعة يمكن أن يكون أكثر تعقيداً عند استخدام ثنائيي اللغة من تنفيذ تصميم مجموعتين متعادلتين. إن أكثر الخيارات المباشرة هي جعل مجموعة واحدة تخضع لصيغة الاختبار A، بينما تخضع الأخرى لصيغة الاختبار B بالرغم من أن هذا الخيار يوازي حالة المجموعتين المتعادلتين (تخضع كل مجموعة لاختبار لم يتم التطرق إليه أو صيغة معتمدة)، فإنه ليس الأمثل عند اختبار ثنائيي اللغة. عند استخدام هذا التصميم فإنه لا يمكن تقييم أداء المجموعة الأولى في اللغة B، ولا أداء المجموعة الثانية في اللغة A يكون الخيار الأفضل هو جعل كل مجموعة تخضع لصيغة مختلطة تحوي على أسئلة من كلا اللغتين، اللغة A واللغة B.

قام سيرسي وبيربيرولو (2000) بإعطاء مثال عن هذا النوع من التصميم المقدم للغة المختلطة. وقد قيموا دقة الترجمة لكلا المجموعتين من الأسئلة من الصيغتين من نموذج تقييم المدرس. إن النسخة الأصلية لهذا الاختبار كانت باللغة التركية بينما كانت الصيغة المكيفة بالإنكليزية. للسيطرة على أثر الممارسة، أجاب المتحنون على صيغة لغوية واحدة فقط من كل سؤال. ولكن ظهرت الأسئلة الإنكليزية و التركية على كل من نموذجي الاختبار. وقد تم ذلك بالتبديل بين اللغتين في كل نموذج. في النموذج الأول، كانت كل الأسئلة الفردية الرقم بالإنكليزية وكل الأسئلة الزوجية الرقم بالتركية. ظهر المخطط المعكوس على الصيغة الثانية. مثلاً، إن السؤال رقم واحد على النموذج الأول ظهر بالإنكليزية بينما ظهر نظيره التركي كالسؤال الأول على النموذج الثاني. السؤال الثاني على النموذج الأول كان بالتركية ونظيره الإنكليزي ظهر كالسؤال الثاني على النموذج الثاني، وهكذا دواليك. علاوة على ذلك، تم إدراج سؤالين إنكليزيين على كل نموذج. قدمت هذه الأسئلة معياراً تم استعماله ضمن تحليل IRT للبرهان على أن فرضية المجموعات المتعادلة بشكل عشوائي صحيحة. تم توضيح دراسة هذا التصميم في الشكل 1.5 . تم استعمال الأسئلة المقدمة بالإنكليزية في كلا النموذجين (معيار 1 و 2) لتقدير إذا ما كان المتحنون ثنائيو اللغة الذين يخضعون لكل نموذج اختباري متعادلين بشكل عشوائي. تم استعمال تحليلات IRT المرتكزة على DIF لاختبار إذا ما كانت الصيغتان الإنكليزية والتركية لكل سؤال يمكن أن تثبتا باستعمال وحدات القياس نفسها.

الصيغة الثنائية اللغة ١

المعتمد ١ (إنكليزي)	المعتمد ٢ (إنكليزي)	السؤال ١ (إنكليزي)	السؤال ٢ (تركي)	السؤال ٣ (إنكليزي)	السؤال ٤ (تركي)

الصيغة الثنائية اللغة ٢

المعتمد ١ (إنكليزي)	المعتمد ٢ (إنكليزي)	السؤال ١ (تركي)	السؤال ٢ (إنكليزي)	السؤال ٣ (تركي)	السؤال ٤ (إنكليزي)

الشكل رقم 1.5 مثال عن تصميم إدارة اللغة المختلطة للممتحنين ثنائيي اللغة.

يعد أثر التبديل بين اللغات على أداء الممتحنين غير معروف. توجد استراتيجية بديلة وهي الحصول على قسمين منفصلين من الاختبار لكل لغة. نصح سيرسي وبيربيرولو (2000) بإجراء المقابلات مع الممتحنين ثنائيي اللغة لاكتشاف إذا كان تغيير لغة الأسئلة ضمن الاختبار شيئاً مربكاً أو معيقاً لأدائهم بطريقة ما.

وقد استنتج سيرسي وبيربيرولو (2000) أن ثنائيي اللغة مفيدون لتحديد الأسئلة التي لم تكن متعادلة عبر اللغات. من ناحية ثانية، لقد صرحوا أن هذا الإجراء لا يمكن أن "يثبت" أن الأسئلة غير المصنفة DIF كانت متعادلة عبر اللغات. لكنهم أفادوا أن الأسئلة التي لم تظهر DIF هي المرشحة بشكل أقوى لتثبيت المعايير اللغوية المنفصلة من الأسئلة المصنفة أو التي لم يتم تقييمها.

كان هناك قصور في دراسة سيرسي وبيربيرولو (2000) وهو أنه تم استعمال نمط واحد فقط من ثنائيي اللغة الخاضعين للاختبار. تضمنت عينة ثنائيي اللغة طلاباً في الجامعة التركية حيث كانت الإنكليزية هي اللغة الأساسية للتعليم. على الرغم من أنهم لم يقوموا بتصفية الطلاب الذين ذكروا بتقاريرهم أنهم "قليلو البراعة" في قراءة أو فهم الإنكليزية، فإن تصميمهم لم يشمل على أي من ثنائيي اللغة الذين كانت الإنكليزية لغتهم الأولى. إن تقييماً أكثر شمولاً لأمانة الترجمة سيحوي كل من ثنائيي اللغة الإنكليزية - التركية و ثنائي اللغة التركية - الإنكليزية. وهكذا، في حال أمكن ذلك، فإنه يمكن تطوير تصميم مجموعتي ثنائيي اللغة بإدراج أكثر من نمط واحد من الممتحنين ثنائيي اللغة.

تصميم رباعي - المجموعات: هناك إضافة واضحة على تصميم المجموعتين ثنائية اللغة وهي جعل مجموعتين ثنائية اللغة تختلف من حيث اللغة الأصلية، للخضوع إلى كلاً من صيغتي الاختبار. تشمل المجموعة الأولى ثنائيي اللغة المتمكنين من اللغة الأولى، بينما تشمل المجموعة الثانية ثنائيي اللغة المتمكنين من اللغة الثانية. يتم تعيين الأفراد في كل مجموعة لواحدة من صيغتي الاختبار (قد تكون

لغة مختلطة). بالإضافة إلى تأمين مجموعات ممثلة أكثر لثنائيي اللغة، فإن هذا التصميم يسمح بتحليل الفروقات الأداء بين النمطين من ثنائيي اللغة. إن تحليلات DIF يمكن إجراؤها بشكل منفصل لكل مجموعة. مثلاً، إذا بدا أن سؤالاً ما متعادل إحصائياً لكل من المجموعتين التركية - الإنكليزية و الإنكليزية - التركية ثنائيتي اللغة، فإنه يتم جمع حقائق أخرى تفيد بأن الأسئلة هي "نفسها" في كلتا اللغتين. إذا أظهر أن سؤال ما DIF في واحدة من المجموعات ثنائية اللغة دون أن يظهر في المجموعة الثانية، تجمع معلومات متعلقة بتفسيرات الاختلافات اللغوية للسؤال.

إذا تم استخدام صيغتي اختبار أحادية اللغة في تصميم رباعي - المجموعات، فإن التحليل التقليدي للإجراءات المتباينة يمكن أن يفيد في تقييم تأثيرات الاختبار "الترجمة" وتأثيرات المجموعة. تم تصوير هذه الحالة في الشكل 2.5، الذي يحوي تحليلاً لتقييم صيغتين افتراضيتين إنكليزية وإسبانية.

نموذج اختباري

اللغة المهيمنة	إنكليزي	إسباني
إنكليزي	؟	؟
إسباني	؟	؟

<u>تفسيرات ونتائج محتملة:</u>

-لا توجد تأثيرات: تدعم تكافؤ صيغ الاختبار عبر اللغات

-التأثير الرئيس لصيغة الاختبار: مشكلة ترجمة

-التأثير الرئيس للغة المهيمنة: اختلاف المجموعة

-التأثير التفاعلي: المجموعة و/ أو صيغة الاختبار غير متكافئة، لا يوجددعم لتعادل الترجمة

الشكل 2.5 تصميم افتراضي رباعي - المجموعة

إذا وُجد تأثير تفاعلي، قد يشير إلى أن اللغة متى تم اختبارها فإنها تشكل فرقاً بحسب المجموعة التي جرى عليها الاختبار. إذا وُجد التأثير الرئيس لصيغة الاختبار اللغوية، فسيشير إلى: مشكلة تتعلق بالترجمة. إن التأثير الرئيس للغة المهيمنة للمجموعة سيشير إلى أن النمطين من ثنائيي اللغة غير متساويين. يمكن إجراء عدة تحليلات باستخدام تغيرات تابعة مختلفة (مثلاً: علامات السؤال، أو علامات الاختبار الإجمالية، أو الدرجات الثانوية على مجالات ضمنية محددة). وهكذا، فإن الإضافات على التصميم الثنائي - المجموعة سيقدم معلومات متزايدة تتعلق بالتفاعل بين اتجاه اللغة الأصلية لثنائيي اللغة واللغة الأصلية للسؤال.

تصميمات متعددة - المجموعات: يمكن للتصميمات رباعية - المجموعات. أن تُوسع بشكل طبيعي لمجموعات أكبر. على سبيل المثال، قد يشمل تصميماً ما لمجموعة من الممتحنين الذين يعتبرون" متمكنين بشكل متساو" من كلا اللغتين. ويمكن استخدام التصاميم أيضاً للتعامل مع ثنائيي اللغة الذين ينتمون إلى خلفيات متعددة بشكل منفصل. مثلاً، قد ترغب دراسة تقارن بين الصيغتين الإنكليزية والإسبانية بالاطلاع على الفروقات بين الكاريبيين، والأمريكيين المتوسطين (أمريكا الوسطى) والمكسيكيين و ثنائيي اللغة الإسبانية - الإنكليزية في جنوب أمريكا. إن الخيار المحدد لعدد المجموعات في التحليل يجب أن يُحرك بواسطة اهتمامات تتعلق بتصاميم بحثية تقليدية كتحديد المتغيرات الخارجية وحجم العينة وقياس الفرضيات المنافسة المعقولة.

باعتبار أن عدد المجموعات المحتملة يتزايد، يظهر سؤال بديهي، وهو هل يمكن للمقاييس المتواصلة لهيمنة اللغة أن تندمج في التصميم البحثي بدلاً من استخدام مجموعات متعددة غير مترابطة. على سبيل المثال، مقاييس كفاءة اللغة A واللغة B يمكن أن ترتد عبر السؤال وبيانات أداء الاختبار لاكتشاف إذا ما كانوا مرتبطين بفهارس DIF فروقات علامات الاختبار الإجمالي. مثلاً، بينوك - رومان (1995) استخدم التحليل الارتدادي لتحديد آثار العوامل اللغوية على أداء

اختبار (GRE اختبار تقرير التخرج) لثنائيي اللغة الإسبانية - الإنكليزية بيرتو ريكان. كان التركيز الرئيس لتحليلها حول إذا ما كانت الصيغة اللغوية لآثار اختبار تؤثر على الاستنتاجات المكتسبة حول المتحنين ثنائيي اللغة. وقد وجدت باستخدام هذه التصميمات أن الكفاءة باللغة الإنكليزية فسرت تباين علامات الاختبار الشفوي GRE حتى 34 ٪. على الرغم من أن تحليلها لم يتحقق من اختبارات المقارنة المقدم بلغات مختلفة، فهي موضحة لأنواع المعلومات التي يمكن جمعها باستخدام التصاميم المتطورة للمجموعات الثنائية اللغة.

تصميمات تستخدم ثنائي اللغة وأحادي اللغة:

تم مناقشة قصور التصميمات التي تستخدم مجموعات منفصلة من المتحنين أحاديي اللغة سابقاً باعتباره قصوراً متعلقاً بالتصاميم الثنائية اللغة. إن التصميمات الأحادية اللغة محدودة لأن هذه النماذج غير قادرة على تحقيق تعادل قياسي كامل. أما التصميمات الثنائية اللغة فهي محدودة بسبب المشكلة التمثيلية. يقترح هذا القسم تصميماً أكثر شمولاً يستخدم كلا النمطين من المتحنين.

استخدام ثنائيي اللغة لتحديد أسئلة معتمدة (معيارية) للتحليلات أحادية اللغة.

تحتاج التصاميم أحادية اللغة إلى بعض الآلية لتفسير الاختلافات في المهارة بين المجموعتين اللغويتين. لو توفر معيار خارجي مرتبط بقوة بالمهارة التي يتم قياسها، فمن الممكن أن يستخدم لتعديل اختلافات المجموعة في الاختبارات. ولكن المعايير الخارجية الصالحة نادرة، هذا إذا توفرت على الإطلاق. خيار ثان لتفسير الاختلافات اللغوية للمجموعة، هو استخدام مجموعة أسئلة متعادلة بما يتعلق بقياس سرعة ودقة العمليات العقلية في اللغتين (أسئلة معتمَدة). بتقديم أسئلة متعادلة، يمكن استخدام الفروقات في أداء المجموعة في هذه الأسئلة، وذلك لتعديل الدرجات في واحد أو أكثر من صيغتي الاختبار (كما تم في تحليلات التعادل التقليدية). أو يمكن استخدام هذه الأسئلة لمعايرة الأسئلة الأخرى على مقياس شائع

كما سنشرح لاحقاً (انظر إلى وودوكومونوز- ساندوفال، 1993، لتوضيح هذه العملية).

إن احتمال كون ثنائي اللغة مفيدون بشكل خاص يكمن في تحديد الأسئلة المتعادلة. إذا تم تقييم مجموعة من الأسئلة لـ DIF عبر اللغات باستخدام تصميم شامل لمجموعة ثنائية اللغة (مثل تصميم رباعي - المجموعات المشروح سابقاً)، فإنه يمكن استخدام الأسئلة التي لا تظهر DIF وذلك لتطوير مجموعة من الأسئلة (المعيارية) المعتمدة التي يمكن استخدامها لربط صيغ لغوية متعددة لأسئلة أخرى بمقياس عام. مثلاً، استخدام طرق قياس (IRT) حيث يمكن قياس صيغتين لغويتين بشكل منفصل في الوقت نفسه (باستخدام ممتحنين أُحاديي اللغة)، ووحدات القياس للأسئلة المعتمَدة (معرّفة من حيث استخدام ثنائيي اللغة) يمكن أن يتم إجباره ليصبح متساوياً عبر المجموعتين اللغويتين. إن نتيجة هذه القيود هي إحداث مقياس عام لكل أسئلة الاختبار الأخرى (مفترضين بالطبع أن الافتراض اللا بعدي لـ IRT يتعلق بالبيانات، وأن مجموعة الأسئلة المعتمدة تمثل على نحو كاف المنشأ الذي تم قياسه). هناك بديل للطريقة المرتكزة على IRT وهي معايرة النموذجين اللغويين بشكل منفصل ثم استخدام الأسئلة المعتمَدة (المعيارية) لتعديل وحدات القياس من صيغة ما إلى مقياس الصيغة الأخرى. (مثلاً، انظر إلى الستوكينغ ولورد، 1983، الطريقة التحويلية). إن تعادل هذه الأسئلة، بالطبع، سيحتاج أيضاً أن يُحفظ على أساس التحليلات النوعية التي أجريت من قبل اختصاصي الاختبارات الثنائية اللغة. إن مجموعة الأسئلة المعتمَدة ستستلزم أن تكون ممثلة للاختبار بأكمله من حيث المميزات الإحصائية والضمنية.

هل تؤكد التحليلات الثنائية اللغة أن الأسئلة المنتقاة على أنها معتمدة هي فعلاً متعادلة عبر اللغتين؟ لسوء الحظ، هي ليست كذلك. إلا أنه باقترانها مع تطوير اختبار "الحالة- الفنية" وطرق تكييف الاختبار (هامبلتون، 1994، موليس وكيلي

وهالي، 1996، انظر أيضاً هامبلتون، المقطع 1، هذا المجلد)، يمكن إعطاء الكثير من الإثباتات لدعم استخدام هذه الأسئلة كأسئلة معتمدة. مثلاً، إن تصميماً ثنائي اللغة لمجموعة رباعية مقترناً بتطور اختبار صائب وطرق تكييفية يمكن أن تزودنا بالأنماط التالية من الدلائل المتعلقة للتعادل المختص بقياس سرعة ودقة العمليات العقلية للأسئلة المنتقاة على أنها معتمَدة:

- يعتبر معدو الاختبار أنه على الأسئلة أن تقيس التراكيب نفسها في كلا اللغتين.
- يُعتقد أن الأسئلة متعادلة من قِبل خبراء موضوعات البحث (مثلاً، علماء النفس أو خبراء المقررات التعليمية).
- يُعتقد أن الأسئلة متعادلة من قِبل الخبراء اللغويين.
- لا تظهر الأسئلة DIF لثنائيي اللغة الذين كانت لغتهم الأم لغة المصدر.
- لا تظهر الأسئلة DIF لثنائيي اللغة الذين كانت لغتهم الأم اللغة الهدف.

علاوة على ذلك، في حال توفر البيانات المعيارية المستقلة، فإن العلاقات الإحصائية بين الأسئلة والمعايير الخارجية يمكن أن تُدرس للتيقن من أن هذه العلاقات متشابهة عبر اللغات.

على الرغم من أن هذه الأنماط المتغايرة من الدلائل لا تؤكد أن الأسئلة هي نفسها في كلتا اللغتين، إلا أنها بالتأكيد تقدم برهاناً قوياً أن الأسئلة التي تستوفي هذه الشروط مناسبة لتكون أسئلة معتمَدة. من الواضح أن استخدام أسئلة كهذه كأسئلة معتمَدة يبرهن من خلال التصاميم السابقة أن معيار الأسئلة اللغوية المختلفة بشكل متوافق بدون استخدام سؤال معتمَد أو جعل التحولات المعيارية مرتكزة على أسئلة تستوفي عدداً قليلاً فقط من هذه المعايير.

معالجات تحليل البيانات

في القسم السابق أُشير إلى أن الأسئلة التي حُددت على أنها متعادلة بما يتعلق بقياس سرعة ودقة العمليات العقلية يمكن أن تستخدم للمساعدة في تشكيل معيار عام أو مشترك عبر نموذجي اختبار بلغتين مختلفتين. وفي هذا القسم، يتم شرح عدة خيارات لإنجاز هذه المعايرة.

افترض أن هذه الحالة اعترضت مختصاً معنياً بقياس سرعة ودقة العمليات العقلية الذي أكمل سلسلة من الدراسات الشاملة المتعلقة بالـ DIF باستخدام ثنائي اللغة. بناء على المقياس الإحصائي والحاسم المنصوص سابقاً، فإن هذا المختص قد حدد مجموعة من الأسئلة ليتم استخدامها كأسئلة معتمَدة. وقد قام هذا المختص بتحصيل بيانات عن كلا الصيغتين اللغويتين للاختبار من المجموعتين المعنيتين أحاديتي اللغة. يتوفر لدينا هنا خياران. الأول، يمكن للمختص المعني بقياس سرعة ودقة العمليات العقلية أن يعاير في الوقت نفسه الأسئلة اللغوية المختلفة على مقياس عام؛ وذلك بإجبار وحدات القياس للأسئلة المعتمدة أن تكون متعادلة عبر المجموعتين اللغويتين. سيتم تقدير القياس للأسئلة الأخرى بشكل منفصل لكل صيغة لغوية للسؤال. السؤال الثاني. هو أن يتم القياس بالمزيد من تحليلات DIF، وذلك باستعمال الأسئلة المعتمدة لتشكيل مقياس عام عبر المجموعتين اللغويتين. بعد أن يتم تحديد الأسئلة التي تعمل بشكل مختلف عبر اللغتين، يمكن أن تتم معايرة الاختبار وذلك بتوجيه كل الأسئلة الأخرى (مثلاً، الأسئلة التي ليست DIF) إلى أن تكون متعادلة عبر اللغتين. إن المعالجة التفاعلية تستخدم كلاً من الممتحنين ثنائيي اللغة وأحاديي اللغة وذلك لربط الفحوص اللغوية المختلفة إلى مقياس عام.

بافتراض هذين البديلين، وبسبب نقص أساس بحثي قوي للاختيار بينهما، فإن الخيار الثاني يبدو أفضل كونه يقوم بالمزيد بالتحليلات DIF باستخدام طريقة IRT، فإن هذه الطريقة ستقوم بتقدير وحدات القياس بشكل منفصل للمجموعتين

اللغويتين المختلفتين لكل الأسئلة ما عدا الأسئلة المعتمدة (التي توجه وحدات قياسها إلى أن تكون متساوية).

عند ذلك يمكن أن يتم تقدير هذه الأسئلة لـ DIF باستعمال طريقة نسبية راجحة (مثالاً، سيرسي بيربيرولو، 2000، سيسن، ستنبيرغ، ووينر، 1988، 1993) إلا أن الطرق التي لا ترتكز على IRT، DIF يمكن أيضاً أن تطبق باستعمال أداء الممتحنين في الأسئلة المعتمدة كالمتغيرات المتطابقة (مثلاً: آلالوف وآل، 1999، بادجل وآل، 1995، سيرسي وآلالوف، 2003) إن المعايرة النهائية المتواصلة سوف تقدر وحدات القياس لأي أسئلة لا تظهر DIF في التحليلات السابقة بشكل منفصل لكل مجموعة لغوية، وسوف توجه وحدات القياس لتكون متساوية لتلك الأسئلة التي لم تظهر DIF آنغوف وكوك، 1988على الرغم من أن طرق معالجة نظرية الاختبار الكلاسيكية يمكن أن تستخدم للمعايرة النهائية، فإن معايرة IRT هي المرجحة (افتراض المدركات اللا بعدية)، بتقديم مميزاتها الإحصائية هامبلتون وآل، (1991).

على الرغم من أن هذه الفكرة مغرية نظرياً، فإن تطبيقات هذه الطريقة سوف تساعد بتحديد فائدتها. إنه لمن المهم أن نلاحظ أنه بغض النظر عن الاستراتيجية التحليلية البيانية المنتقاة، فإن صدق مجموعة الأسئلة المعتمدة هو شيء حاسم. يفترض التحليل الموجز هنا أن مجموعة الأسئلة المستخدمة لاعتماد المقياس عبر اللغات مناسب. إن صدق هذا المعتمد يجب أن يتم دعمه باستخدام معايير خارجية كأحكام خبراء موضوعات البحث وتحليلات تعادل التركيب سيرسي وآل، (1999) إن الأسئلة المعتمدة يجب أن يتم اعتبارها كممثلة لأشكال الاختبار الكامل من حيث المميزات الإحصائية والضمنية.

هناك خيار آخر لم يتم تطبيقه بعد على مجال تقييم التقاطعات اللغوية، ألا وهو استعمال الأسئلة المعتمدة كواحدة من المقاييس العدة لمطابقة الخاضعين للاختبار بلغات مختلفة. مثلاً، إن الممتحنين في مجموعات لغوية مختلفة يمكن أن

تتم مطابقتهم على معيار المطابقة المتعددة التغيرات التي تشكل الأداء في الأسئلة المعتمدة، درجات المناهج المعنية، الحالة الاجتماعية الاقتصادية، والمعتقدات المتغيرة الأخرى المرتبطة بالتركيب. أشار سيرسي (1997) أن ميل الدرجات يمكن أن يستخدم لمزاوجة الممتحنين في هذا السياق. بالإضافة إلى ذلك تم إجراء دراسات DIF عبر المجموعات التي تجيد اللغة نفسها (مثلاً. الإناث والذكور) باستخدام المنطق الرمزي للتراجع في سلوك الفرد: وذلك لتكييف التحليلات باستخدام متغيرات متعددة (كلوزر، نانجستر، مازور وريبكي، 1996، مازر، كانجي وكلورزر 1995). تحوي هذه الإستراتيجية وعداً لربط مقاييس الدرجات عبر الصيغ اللغوية المختلفة لاختبار ما.

الاستنتاجات

في هذا المقطع، تم تنقيح استعمال التصاميم البحثية المتضمنة ممتحنين ثنائيي اللغة وذلك لتقدير الاختبارات المقدمة بعدة لغات. إن التقنيات المتعلقة بقياس وسرعة ودقة العمليات العقلية في هذا المجال هي فقط بطور التطور: لذلك هناك حاجة للمزيد من البحوث التجريبية. بالطبع، قد لا يتوفر ممتحنون ثنائيو اللغة في كل حالات تقييم التقاطع اللغوي. ولكن في تلك الحالات حيث يمكن إدراج ثنائيي اللغة في التصميم البحثي يمكن جمع دلائل أكبر تتعلق بالاختبار ومقارنة الأسئلة. بافتراض الدروس التي تم تعلمها في هذا التنقيح، تقدم بعض الاقتراحات باستخدام ثنائيي اللغة لتحسين تعادل الاختبار عبر اللغات.

أولاً، إن دراسات تعادل الاختبار عبر اللغات يجب أن يتضمن تحليلات لكل من الممتحنين الثنائيي اللغة والأحاديي اللغة. تعد التصميمات التي تستخدم ثنائيي اللغة مفيدة بشكل استثنائي لتقييم تعادل الأسئلة عبر مجموعة عادية للخاضعين للاختبار. يجب أن تقدم نتائج هذه التحليلات دلائل قيمة لاختيار الأسئلة المعتمدة ليتم استخدامها في التحليلات اللاحقة. ولكن يجب ألا تتخذ القرارات المتعلقة

بالإدارة والمجموعة وتطور الاختبار فقط على أساس التحليلات التي تستخدم ثنائيي اللغة. بل إن هذه التحليلات يجب أن تكون جزءاً من دراسة أكثر شمولاً والتي تقيّم بدورها أداء المجموعات الأحادية اللغة في كل صيغة لغوية من الاختبار، ويجب أن تتحرى عن العلاقات بين نقاط الاختبار ونقاط الأسئلة ومتغيرات أخرى ضمن شبكة علم القوانين الطبيعية والمنطقية المتعلقة بالمنشأ المقاس.

ثانياً، حين يتم استخدام ثنائيي اللغة لتقدير الاختبار وتعادل الأسئلة عبر اللغات، فإنه يجب ألا تعامل المجموعات الثنائية اللغة كمجموعة واحدة. كحد أدنى، إن أداء المجموعتين ثنائيتي اللغة الذي يمثل الهيمنة في كل من اللغتين يجب أن يتم التحري عنها. لهذا، هناك ميزة رئيسة لتصاميم البحث الثنائية اللغة ألا وهو الآلية لتصنيف ثنائيي اللغة على مجموعتين أو أكثر، بالإضافة إلى التصديق على أنهما متمكنتين من كلتا اللغتين (بيكر 1988). لا يوصى بتصاميم الثنائية اللغة أحادية المجموعة حيث يخضع ثنائيو اللغة إلى كلتا الصيغتين اللغويتين للاختبار أو الأسئلة، ويعود ذلك إلى مشكلات تتعلق بالتعب والحافز وأثر الممارسة.

ثالثاً، إن فوائد استخدام صيغ الاختبار المختلطة اللغة بدلاً من صيغ الاختبار الأحادية اللغة يجب أن تؤخذ بعين الاعتبار في التصاميم الثنائية اللغة. تقوم الصيغ اللغوية المختلطة بجمع بيانات عن اللغتين من كلتا المجموعتين ثنائيتي اللغة. تعتبر هذه الطريقة نافعة لأنها تمنح لكل الممتحنين الفرصة لإثبات مهارتهم في مجال الموضوع الذي يتم اختباره في كلتا اللغتين. ولكن إذا أولينا الأهمية الكبرى للآثار التفاعلية للّغة الأصلية في لغة الاختبار، فإن التصميم الذي يستخدم صيغاً لغوية منفصلة يمكن أن يكون هو الراجح. في كلتا الحالتين، يجب أن تكون المجموعات التي تخضع لصيغ الاختبار متعادلة بشكل عشوائي (عبر الفروض العشوائية أو أشكال الاختبار الحلزوني). علاوة على ذلك، فإن افتراض التعادل العشوائي يمكن أن يتم قياسه باستخدام عدة أسئلة شائعة (يفضل في كلتا اللغتين) في كلتا الصيغتين.

رابعاً، إن أثر الـ DIF عبر اللغات يجب أن يتم تقديره مع الأخذ بعين الاعتبار مجالات المضمون المتعددة المدرجة في التقدير. إذا كانت أسئلة الـ DIF مترافقة بشكل سائد مع بعض مجالات المضمون، فإن المقارنات عبر المجموعات اللغوية فيما يتعلق بمجالات المضمون قد لا يمكن تبريرها. بالنظر إلى نماذج الـ DIF ضمن مجالات المضمون سيظهر قصور على أنماط استنتاجات التقاطعات اللغوية التي يمكن تطبيقها.

كانت مقاييس الاختبارات النفسية والتربوية (جمعية البحوث التربوية الأمريكية [AERA]، الجمعية النفسية الأمريكية [ABA]، والمجلس الوطني للقياس في التعليم [NCME]، 1985) واضحة في طلب حقائق عن مقارنة الاختبارات المقدمة بعدة لغات: "حين يكون القصد أن صيغتي اختبارات اللغة المزدوجة قابلة للمقارنة، فإنه يجب أن يورد إثبات مقارنة الاختبار ص.75" . تم تأكيد هذا النموذج بشكل خاص في التنقيح الحديث لهذه النماذج:" حين يقصد من الصيغ اللغوية المتعددة لاختبار ما أن يكون قابلاً للمقارنة فيجب على مطوري الاختبار أن يقوموا بالتبليغ عن إثبات مقارنة الاختبار" (AERA، ABA، NCME، 1999، ص99). إن التغيير من "صيغتين" إلى صيغ "متعددة" يجب أن يسلم بالتزايد الكبير في تقييم التعدد اللغوي خلال الـ 15 السنة الماضية منذ أن نشرت الطبعة الأخيرة لمقاييس الاختبار. إن الاقتراحات الموجزة في هذا المقطع يجب أن تساعد الباحثين مطوري الاختبارات ببذل أفضل ما يمكن في تقييم الصيغ اللغوية المختلفة لاختبار ما وفي تقديم شاهد لمقارنة الدرجات.

إن بعض الاختبارات، مثل المسابقات الرياضية البارزة في عالم الأولمبياد، تتجاوز الحواجز اللغوية. لسوء الحظ، إن مقارنات المعرفة والمهارات النفسية الأخرى هي إجمالاً لا تقاس باستخدام التقييمات "المستقلة لغوياً". هناك على الأغلب عوامل كثيرة تتعلق بتقييم التقاطع اللغوي للاستنتاج المطلق. إن اختلافات

الاختبار يمكن أن تكون منفصلة بشكل كامل عن اختلافات المجموعات اللغوية. لذا، يجب أن نبذل كل ما بوسعنا لتفسير التأثيرات اللغوية حين نقوم بالمقارنات للأفراد الذين يتكلمون لغات أخرى. إن دراسة أداء الاختبار لثنائيي اللغة في اختبارات اللغة المزدوجة تقدم إطاراً واعداً لتقييم هذه التأثيرات.

*****

المراجع

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55–73.

Allalouf, A., Hambleton, R K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36,* 185–198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and*

psychological testing. Washington, DC: American Educational Research Association.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Report No. 88-2). New York: College Entrance Examination Board.

Baker, C. (1988). Normative testing and bilingual populations. *Journal of Multilingual and Multicultural Development, 9,* 399–409.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli Universities. *Educational Measurement: Issues and Practice, 13,* 12–20.

Boldt, R. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers* (College Entrance Examination Board Research and Development Report 68-69, No. 3). Princeton, NJ: Educational Testing Service.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202–214.

CTB/McGraw-Hill. (1988). *Spanish assessment of basic education: Technical report.* Monterey, CA: McGraw-Hill.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74,* 912–920.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177–184.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3&4), 199–215.

Foster, D., Olsen, J. B., Ford, J., & Sireci, S. G. (1997, March). *Administering computerized certification exams in multiple languages: Lessons learned from the international marketplace.* Paper presented at the meeting of the American Educational Research Association, Chicago.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6.* 304–312.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57–68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross -national ssurveys of educational achievement* (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests [Special Issue]. *Language Testing, 20*(2), 127–240.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–157.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guideleines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1–16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818–825.

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology, 71,* 83–94.

International Association for the Evaluation of Educational Achievement. (1994). *TIMSS main study manuals: Population 1 and 2.* Hamburg, Germany: Author.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.

Mazor, K. M., Kanjee, A., & Clauser, B. (1995). Using logistic regression and the Mantel–Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131–144.

Mislevy, R. J., & Bock, R. D. (1990). *PC-BILOG 3: Item analysis and test scoring with binary logistic items.* Mooresville, IN: Scientific Software.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation verification procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection* (pp. 1–14). Chestnut Hill, MA: Boston College.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Pennock-Román, M. (1995). *Measuring developed academic abilities using Spanish- versus English-language tests: PAEG/GRE relationships for Puerto Ricans who are more proficient in Spanish than in English* (GRE Report No. 89-01). Princeton, NJ: Educational Testing Service.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*(1), 12–19, 29.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 147–165.

Sireci, S. G., & Berberolu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education, 13*(3), 229–248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Sireci, S.G., Xing, D., & Fitzgerald, C. (1999, April). *Evaluating translation DIF across multiple groups: Lessons learned from the Information Technology industry.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Terman, L. M. (1916). *The measurement of intelligence.* Boston: Houghton-Mifflin.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum Associates.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Norwood, NJ: Ablex.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1977). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29–37.

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47*(4), 263–279.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment, 3,* 1–16.

6

إرساء قواعد مقارنة الدرجات لاختبارات معطاة بلغات مختلفة

ليندا ل. كوك

خدمة الاختبارات التربوية

اليسا ب. شميت - كاسكالار

مجموعة التقييم العالمية، بروكسل، بلجيكا

إن الاختبارات المكيفة في الإدارة لمجموعات لغوية مختلفة وإعطاء الاختبارات المكيّفة للمتقدمين للاختبار من ثقافات مختلفة هو تدريب يملك تاريخاً طويلاً في مجال التقييم النفسي. يشير عمل "نيرمان" (1916) إلى كم مضى من الزمن على معرفة الباحثين بالمشكلات المتصلة باستخدام الأدوات التي كانت قد طوّرت من أجل سكان بلد ما لتقييم صفات سكان بلد ثان والذي ربما يختلف عن الأول في الخلفية والثقافة.

إن استعمال اختبارات مكيّفة تم تطويرها لسكان بلد معين، ومن ثم إعطاء هذه الاختبارات إلى سكان بلد ثان، والذي قد يختلف عن الأول في كل من اللغة والثقافة، هو تدريب ازداد بدرجة كبيرة على مدى العشر سنوات الماضية. وضع هامبلتون (1993) وسيرغي (1997) في قائمة عدداً من العوامل المساهمة في

الاهتمام المتزايد في تكييفات الاختبار. ومن بين تلك العوامل: تعزيز العدالة في مقارنات الأفراد والمجموعات من خلفيات ثقافية ولغوية مختلفة: تسهيل الدراسات المقارنة عبر المجموعات الثقافية والعرقية والوطنية؛ تسهيل مقارنة إنجاز الطلاب في بلدان مختلفة. إضافة إلى هذه القائمة تأتي العولمة المتزايدة لأعمال تجارية عديدة، مؤدية الحاجة لتطوير وتكييف اختبارات في اللغة الأصلية لموظفين من أجل استخدامها في المصادقة عليها رسمياً. إن كلاً من العوامل مستقل في القدرة على مقارنة الدرجات النهائية المحصّلة من الاختبارات المقررة على مجموعات تختلف في كل من اللغة والثقافة.

بغض النظر عن الأسباب لأجل تكييف اختبار جرى إعطاؤه في إحدى اللغات إلى الإدارة في لغة ثانية أو في لغات متعددة، فالموضوعات التي تحيط بعلم المنهج الملائم من أجل تكييف الاختبار لدعم مقارنات ذات مصداقية للدرجات النهائية تكون معقدة إلى حد كبير.

أشار بورتينغا (1989) إلى أنه ربما تكون مقارنات قدرات الأفراد والجماعات مضللة لسببين: يتصل السبب الأول بالصفة المقاسة، وقد أعطى على سبيل المثال عدم جدوى مقارنة طول شخص ما بوزن فرد ثان. ويتصل السبب الثاني بوحدات القياس المستخدمة في المقارنة: على سبيل المثال، لا يستطيع المرء أن يقوم بمقارنة مباشرة لطول شيئين إذا جرى قياس شيء واحد بالإنش والآخر بالسنتيمتر. تبدو تلك كنقاط واضحة عندما يعود المرء إلى الصفات الفيزيولوجية مثل الارتفاع، والوزن، والطول. على أية حال، يصبح الوضع للتو أكثر تعقيداً عندما تمتد المقارنات إلى درجات نهاية محصّلة من تقييمات تربوية ونفسية.

لنأخذ بعين الاعتبار، على سبيل المثال، اختباراً للجبر يحوي شيئاً من مشكلات الكلمة. لننظر أبعد بأن الاختبار قد تم تصميمه باللغة الإنكليزية وجرى وضع درجاته النهائية باستخدام معطيات من سكان ناطقين بالإنكليزية. ثم يترجم

الاختبار إلى الإسبانية ويقرر على مجموعة طلاب ناطقين بالإسبانية. فإذا لم يأخذ الطلاب الناطقون بالإسبانية درجات نهائية جيدة في الاختبار مثل تلك التي أخذها الطلاب الناطقون بالإنكليزية، كيف لنا أن نعرف فيما إذا كانت الفروق في الدرجات النهائية بسبب اختلاف المجموعات في مقدرتهم في الجبر، أو أن يعزى ذلك إلى حقيقة أن الترجمة لمشكلات اللفظ الجبرية إلى اللغة الإسبانية جعلت جوهرياً المشكلات أكثر صعوبة على المتقدمين للاختبار الناطقين بالإسبانية؟

إمكانية أخرى وهو أن الاختبار المقرر في اللغة الإسبانية يتطلب وقت قراءة أكثر منه في الاختبار المقرر في اللغة الإنكليزية، وهكذا يجعله أكثر توفيقاً للسكان الناطقين بالإسبانية. هل ينبغي أن يكون النجاح عاملاً في تقييم المقدرة بالجبر للمجموعة الناطقة بالإسبانية وليس للمجموعة الناطقة بالإنكليزية.

إضافة إلى ذلك، ربما من المحتمل ألا تكون التعليمات للاختبار مترجمة بوضوح ويكون المتقدمون للاختبار الناطقون بالإسبانية مشوشين نظراً للخطط الاستراتيجية لمفتاح أخذ الاختبار، مثلاً فيما إذا كانوا سيعاقبون أم لا من أجل إجابات تخمينية على الأسئلة.

إن قائمة الأسباب للفروق بين الدرجات النهائية المحصّلة في اختبار للجبر معطى لتوه لمجموعات ناطقة بالإنكليزية ومجموعات ناطقة بالإسبانية هي بالتأكيد القاطع ليست كاملة: المقصود بها فقط هو كم يكون صعباً تجنب بناء مصادر تنوع في الدرجات النهائية للاختبار عند مقارنة الدرجات النهائية في الاختبارات المكيّفة.

لقد قام عدد كبير من الباحثين بوصف الإجراءات التي تناولت قضايا التباين في درجات الاختبار غير المتصل بالموضوع وبالتالي ترويج مستوى متزايد لقابلية مقارنة الدرجات في الاختبارات المكيفة.

(انظر كيسنجر، 1994، وسيرغي 1997، وهامبلتون 1993، من أجل نقاشات عميقة لهذه الإجراءات). تشمل الإجراءات الترجمة، الترجمة الارتجاعية للأداة التي ستكيّف، اختبار الدليل وغربلة مفردات الاختبار لتوظيف مفردة مميزة، اختبار ميداني وموزون، تطوير إجراءات الإدارة، وبحث علمي ذو مصداقية.

هذه النقطة الأخيرة مهمة إلى حد كبير لأنه، على الرغم من درجة الانتباه القصوى المبذولة للمسائل المنهجية، قد لا يكون من الممكن الحصول ببساطة على بناء مواز لاختبار تم إعطاؤه إلى سكان بلدان متعددة تختلف في اللغة والثقافة. بالتالي، من المهم للبحث العلمي ذي المصداقية أن يؤخذ به في أي اختبار مكيّف لتأكيد أن المقارنات والتفسيرات الصادقة للدرجات النهائية يجري تدعيمها بالدرجات النهائية للاختبار.

يجري التركيز في هذا الفصل على عامل واحد فقط مؤثر على قابلية مقارنة الدرجات النهائية المحصلة في الاختبارات المكيفة. السبب الثاني لتضليل المقارنات وفقاً لـ "بورتينغا" (1989) هو وحدات قياس غير متكافئة. في الأقسام التالية من هذا الفصل، نزود قاعدة لفهم الطرائق الإحصائية المتوفرة حالياً من أجل تعادل وموازنة الاختبارات النفسية والتربوية، نصف وننقد إجراءات ربط المقاييس المحددة التي تستخدم في دراسات تكييف الاختبار، ونوضح إجراءات ربط منتقاة وقضايا بوصف ونقد دراسات ثلاث جرى القيام بها عبر العشرين سنة الماضية وذلك لربط الدرجات النهائية في اختبار التقييم للمدارس الثانوية (SAT) Scholastic Assessment Test بالدرجات النهائية في الـ (Prueba de Aptitud Academica) (PAA).

نظرة شاملة على طرق الربط:

ناقش لين (1993) حقيقة أن العديد من التقنيات المختلفة متوفر من أجل ربط نتائج الاختبار وأن المصطلح المستخدم لوصف التقنيات لم يستخدم دائماً بصورة واضحة. استمر لين في وصف خمس طرق مختلفة لربط نتائج الاختبار وكيف يؤثر

نمط الربط على المقارنات والتفسيرات. بيّن القصد من أن التداخلات التي تدّعي قابلية تبادل الدرجات النهائية تتطلب طرائق قوية لربط مقاييس الاختبارات. ربما تكتفي أنماط أخرى للتداخلات بأشكال للربط أضعف، لكن أشكال الربط الأضعف تلك هي بطبيعتها تابعة للسّياق، المجموعة والزمن.

في مقالته 1993، وصف لين خمس طرائق لربط الاختبارات النفسية والتربوية. هنا يتم فقط وصف أربع منها. هذه الطرائق الأربع هي: التساوي، المعايرة، التعديل الإحصائي، والتنبؤ.

التساوي (Equating)

احتفظ لين (1993) بمصطلح "متساوي" من أجل الروابط التي تعطي درجات نهائية يمكن استخدامها بالتبادل. حدد الغاية بأن أقوى صورة لمقياس ربط الدرجة النهائية هو التساوي. يرجع لين إلى لورد (1980) وتعريفه "للتساوي" الذي يتطلب أن يستخدم مصطلح "درجات نهائية متساوية" فقط عندما يكون اختيار أي ترجمة أو شكل للاختبار ينبغي أخذه مسألة لا تقدم ولا تؤخر بالنسبة للمتقدم للاختبار ومستخدم درجاته النهائية. من الواضح أنه إذا كانت المقارنات للدرجات النهائية للاختبار تتطلب اختبارات ينبغي اعتبارها قابلة للتبادل (مثل: درجات نهائية على ورق اختبار لـ SAT مقررة في تشرين الأول ودرجات نهائية على ورق اختبار لـ SAT في حزيران) عندئذ يجب أن تستخدم إجراءات التساوي. أشار لين إلى أن المتطلبات لأجل التساوي هي أن أشكال الاختبار يجب أن تقيس التركيب نفسه بدرجات متساوية من المصداقية، وهذا يعني أن الأشكال يجب أن تكون قابلة للتبادل. أشار آخرون (فان دي فيفر وبورتينغا 1991) إلى أنه ليس فقط يجب أن تكون الأشكال قابلة للتبادل من أجل إجراءات دراسة متساوية مناسبة بل يجب أيضاً أن تكون الشروط الفيزيائية لإدارة الاختبار قابلة للمقارنة.

المعايرة (Calibration)

وصف لين (1993) "معايرة" كوسيلة لمقارنة الدرجات النهائية على أوراق اختبار ترضي المتطلبات الأقل تشدداً من المتطلبات لأجل اختبارات متساوية، أعطى التالي

كأمثلة على "المعايرة": اختبارات الربط التي تختلف بدرجة معتبرة في الطول، وبالتالي في المصداقية: اختبارات الربط التي يمكن أن تستخدم لمقارنة طلاب في مستويات متطورة مختلفة (جرت الإشارة إليها في الأدب كدراسات متدرجة عمودية).

وضع لين (1993) في قائمة المتطلبات من أجل المعايرة (مثل: يجب أن تقيس "الاختبارات" التركيب نفسه. لكن يمكن أن تختلف في المصداقية. يمكن أن تختلف أيضاً في المستوى الذي تكون فيه المقاييس أكثر فائدة) (صفحة 90). أشار لين إلى أن المعايرة تعطي وسيلة لمقارنة الدرجات النهائية على أوراق الاختبار التي ترضي متطلبات أقل تشدداً من تلك التي على أوراق اختبار متوازية. على أي حال، يوجد ثمن ينبغي دفعه. يوجد هنالك الإمكانية بأن يكون بالاستطاعة إجراء نماذج مختلفة متعددة من دراسات المعايرة وستعطي كل معايرة الإجابة عن سؤال مختلف.

استشهد لين باتصال شخصي من مسليفي وستوكنغ كما أشار إلى أنه عندما لا تكون الاختبارات X و Y غير ذات مصداقية بدرجة متساوية تستطيع المعايرة التي تحول الدرجات النهائية Y إلى المقياس X أن تجيب عن السؤال "ما تكون قيمة X لأجله تكون الدرجة النهائية X للشخص الأكثر ميلاً؟". أشار المؤلفون إلى أنه من الأرجح أن تعطي المعايرة نفسها إجابات غير صحيحة لأسئلة حول خواص توزيعات الدرجات النهائية للمجموعات المتقدمة لاختبارات X و Y.

التعديل الإحصائي (Statistical Moderation)

يرجع إجراء الربط الثالث الموصوف من قبل لين (1993)، "التعديل الإحصائي"، إلى إجراءات تشمل عادة استخدام الدرجات النهائية في اختبار خارجي لدرجات نهائية وسطية تم الحصول عليها في الاختبارات التي ينبغي أن تقارن. إن التعديل الإحصائي هو وسيلة تقنية يمكن استخدامها لمقارنة الدرجات النهائية في اختبارات التحصيل التي تقيس مناطق مادة مختلفة. على سبيل المثال، ترغب الكليات أحياناً في مقارنة درجات نهائية اكتسبها طلاب خضعوا لاختبارات مختلفة لمادة SAT II.

يجري تطوير نظام متري عام للدرجات النهائية لاختبار مادة SAT II باستخدام الدرجات النهائية للرياضيات والكلمات لـ SATI وذلك كاختبار خارجي في دراسة التعديل الإحصائي. على العكس من طريقتي الربط التي جرت مناقشتهما سابقاً، لم يتطلب التعديل الإحصائي تقييمين سيجري ربطهما لقياس التركيب نفسه. على أية حال، يتطلب هذا الإجراء عند استخدامه في عمل ربط اختبار خارجي عام، ويعتمد نجاح هذا الإجراء بدرجة كبيرة على متانة العلاقة بين الاختبار الخارجي والمقاييس التي ينبغي ربطها.

إن أحد المساوئ الرئيسة لتقنيات التعديل الإحصائي هو أن مكوناته تابعة للسياق والمجموعة، والزمن. بالتالي ربما تتنوع العلاقة التي جرى تأسيسها بين الدرجات النهائية على تقييمين تم تطويرهما باستخدام تقنيات التعديل الإحصائي وذلك وفقاً لمجموعة المتقدمين للاختبار الذي جرى انتقاؤهم للاشتراك في دراسة التعديل.

التنبؤ (Prediction)

إن إجراء الربط الرابع الذي وصفه لين (1993) هو التنبؤ. علق لين أنه طالما يوجد درجة ما للعلاقة بين الإنجاز في أحد التقييمات مع الإنجاز في آخر، يكون ممكناً الربط بين التقييمين من خلال التنبؤ. بالطبع ستعتمد متانة ربط التقييمين على متانة العلاقة بين الدرجات النهائية المحصّلة من التقييمين. إن بعض النقاط الضعيفة في التنبؤ، عند استخدامه كإجراء ربط، تكون في أن تعادلات التنبؤ تعتمد على المجموعة. أيضاً تكون تعادلات التنبؤ وحيدة الاتجاه؛ مما يعني أنه يجب استخدام روابط منفصلة للتنبؤ بالدرجات النهائية لاختبار Y من اختبار X والدرجات النهائية في اختبار X من اختبار Y.

تطبيق أربع طرق لربط درجات نهائية على أوراق اختبار معطاة في لغات مختلفة

من المهم أن نأخذ بعين الاعتبار التأسيس للدرجات النهائية القابلة للمقارنة من أجل الاختبارات التي قد جرى تكييفها وفق لغات مختلفة وتقررت على مرشحين في لغاتهم الأصلية من منظور نطاق الربط الذي قدمه لين (1993).

أولاً، من الواضح أنه من المحال تقريباً ربط الاختبارات التي تم تكييفها مع لغات مختلفة، ومن ثم إعطاء تلك الاختبارات إلى المتقدمين للاختبار في لغاتهم الأصلية واعتبار كون الاختبارات المربوطة متعادلة. السبب في هذا هو أن إعطاء المشكلات المقترنة بالاختبارات التي تم تكييفها من أجل لغة مختلفة ومجموعات ثقافية مختلفة (هامبلتون 1993) ليس من المحتمل أن افتراض الأشكال المتوازية (أشكال متشابهة جداً في المحتوى والصفات الإحصائية)، (مطلوبة لإجراء التساوي، أن يتحقق).

ينبغي أيضاً أن يكون واضحاً أنه من غير المحتمل أن يستطيع شخص ما أن يقول إنها مسألة لا فرق فيها بالنسبة للمتقدم للاختبار، فيما إذا أخذ أو أخذت اختباراً باللّغة الأصلية أو بلغة مترجمة (لورد 1980، بيّن هذه النتيجة كمحصلة لاختبارات درجات نهائية متساوية). ربما إذا كان بالإمكان تكييف اختبار بشكل تام، وكان المتقدمون للاختبار ثنائيي اللغة بشكل تام فإن حالة كهذه يمكن أخذها بالاعتبار. على أية حال، لا توجد أي من تلك الحالات بصورة حرفية في التقويمات الثقافية المتداخلة/ اللغوية المتداخلة.

إن المعاني المتضمنة في محاولة ربط الاختبارات المكيّفة مع لغات مختلفة وإعطاؤها إلى طلاب بلغتهم الأصلية ولغات مترجمة هي واضحة تماماً. من المحال لأية دراسة ربط، بغض النظر عن العناية التي بذلت لتنفيذها، أن تعطي درجات نهائية قابلة للمقارنة في مفهوم الدرجات النهائية المتساوية.

سؤال مهم يجب طرحه وهو فيما إذا كان ممكناً أو ليس ممكناً ربط الدرجات النهائية للاختبارات التي جرى استخدامها لتقويمات لغات متداخلة واعتبار هذه الدرجات معيارية باستخدام معيار لين (1993). وفقاً لـ"لين"، لا يتطلب الربط عن طريق المعايرة أن تكون الاختبارات ذات ثبات متساوٍ، ولكن أن تقيس المفهوم نفسه فقط. في سبيل الإجابة عن سؤال فيما إذا كانت المعايرة ممكنة أم ليست ممكنة، يجب أن يؤخذ بعين الاعتبار طبيعة التقييم. على سبيل المثال من المقبول بدرجة أكبر

كثيراً، أن يقيس اختبار رياضيات جرى إعطاؤه باللغة الإنكليزية لمجموعة ناطقة بالإنكليزية وباللغة الإسبانية لمجموعة ناطقة بالإسبانية المفهوم نفسه للمجموعتين من أن يقيس اختبار لمقدرة لفظية أعطى تحت نفس الظروف للمجموعتين نفس المفهوم. إذا لم تكن إحدى إجراءات التساوي. أو المعايرة قابلة للتطبيق بسبب طبيعة الاختبارات كما يتمنى المتدرب أن يقوم بالربط. فإنه يلجأ إلى إجراءات التساوي. يجب أن يجري تصميم دراسات التعديل الإحصائي بعناية شديدة لتحقيق الحاجات الفريدة لدراسة ربط اللغات المتداخلة. إن السبب في وجوب تصميم الدراسات بعناية هو أن اختباراً خارجياً عاماً ينبغي أن يؤخذ من قبل مجموعات في اللغة الأصلية وفي لغة الهدف المترجم إليها. من الصعب أن نرى كم يكون هذا ممكناً إذا كانت المجموعتان أحاديتي اللغة بالتمام. على أية حال، من المحتمل أن يوجد طريقان كأن يكون ممكناً محاكاة اختبار خارجي مشترك وأن يكون ممكناً تقرير النتائج، لدراسة التعديل. يتمثل أحد الطرق في تكييف اختبار مشترك وفي اللغة الأصلية ولغة الهدف مع عناية كافية وبذلك سيعمل كعامل ربط "عام" بين قياسين عند إعطائه إلى مجموعتي لغة خاصتين. الطريق الثاني بأن يعطى الاختبار العام إما باللغة الأصلية أو بلغة الهدف إلى مجموعة ثنائية اللغة. وبالتالي ربما يؤدى الاختبار كاختبار خارجي مشترك. وقد جرى استخدام كلا هذين الإجراءين مع بعض النجاح في دراسات ربط عبر لغوية.

إن الإجراء الرابع الموصوف من قبل لين (1993) هو التنبؤ. دراسات تنبؤ مماثلة تماماً أنتجت درجات نهائية للاختبار X جرى التنبؤ بها من الدرجات النهائية للاختبار Y. بغية تطوير معادلة التنبؤ، يجب أن يأخذ بعض المتقدمين للاختبار كلا الاختبارين. حذّر لين من أن النتائج لدراسات التنبؤ لديها عدد من التحديدات. إضافة إلى ذلك، حذّر من أن تكون العلاقة التي تم استخدامها لتقرير معادلات التنبؤ هي لمجموعة معينة. ولهذا عواقب معينة من أجل تطبيق هذا النمط من إجراء الربط لعمليات التقييم عبر اللغات وبافتراض، أن يكون الشرط الأساسي لتطوير

معادلات التنبؤ هو وجوب أن تأخذ مجموعة واحدة كلا الاختبارين، فإن ما يترتب على ذلك هو أن دراسة ربط لغات مختلفة مرتكزة على التنبؤ هو أن المجموعة المستخدمة من أجل دراسة لربط يجب أن تكون ثنائية اللغة. إن الضرر في استخدام مجموعة ثنائية اللغة لهذا النمط من الدراسة هو أن النتائج لهذه الدراسة قد لا تعمم على الحالة مركز الاهتمام، وهي الحالة التي تعطى فيها الاختبارات بلغات أصلية ولغات مترجم إليها إلى متقدمين للاختبار ناطقين بلغة واحدة من تلك اللغات.

إن إحدى ميزات دراسات التنبؤ كأساس لربط اختبارين مقدمين بلغات مختلفة هي أن الإجراءات تسمح باستخدام متغيرات معدّلة للغة وبالتالي ربما تقدم إجابة أكثر دقة للسؤال عن مدى جودة إنجاز الطالب (الفرد) إذا تم إعطاؤه الاختبار باللغة الهدف (المترجم إليها).

ربط اختبارات معطاة بلغات مختلفة:

مثالياً، يرغب أولئك المهتمون في ربط التقويمات التي كان قد تم تكييفها مع لغات مختلفة وجرى إعطاؤها إلى متقدمين للاختبار ناطقين بلغة واحدة في لغاتهم الخاصة، يتمنون أن يكونوا قادرين على مقارنة مهارات وقدرات المتقدمين للاختبار بأخذهم تقويمات مختلفة كما لو كانت الدرجات النهائية المحصلة في التقويمات قابلة للتبادل كلياً (متعادلة). على أية حال، وكما جرت الإشارة إليه سابقاً في هذا الفصل، يكون هذا الوضع المثالي صعباً (إن لم يكن محالاً) الحصول عليه لأن المعطيات التي جرى جمعها في دراسات الربط عبر اللغات لم يتم تزويدها جيداً بنماذج متوازنة نموذجية.

قدم سيرسي (1997) نظرة شاملة ممتازة للموضوعات التقنية المتعلقة باختبارات ربط تم استخدامها في تقويمات عبر اللغات. بدأت مراجعته بمناقشة حقيقة أن المتدربين يعتقدون أن ترجمة اختبار ببساطة من إحدى اللغات إلى أخرى هو شرط كاف لتقويم لغات مختلفة. أشار سيرسي، إلى المغالطة في سير المحاكمة

هذه بأنه لا شيء لأن الآثار غير المقصودة للترجمة قد تنتج مواد تختلف في الصعوبة وفي صفات أخرى عبر لغات مختلفة (انظر كسينجر 1994، هامبلتون، 1993، 1996، أولميدو 1981، بريتو 1992).

وفقاً لسيرسي (1997)، إن التصاميم المستخدمة لربط التقويمات المعطاة في لغات مختلفة تقع في فئات ثلاث:

(أ) تصاميم المجموعة أحادية اللغة المنفصلة.

(ب) تصاميم المجموعة ثنائية اللغة.

(ج) تصاميم المجموعة أحادية اللغة المطابقة.

تتضمن مجموعة أحادية اللغة المنفصلة بالضرورة إجراءاً ما لتطوير المواد المتطابقة جزئياً، بينما يحوي التصميمان الآخران كمطلبهم المركزي تطوير المقاربات لتتطابق جزئياً مع مجموعات المتقدمين للاختبار.

تصاميم مجموعة أحادية اللغة المنفصلة:

تصل كل هذه التصاميم إدارة الاختبارات في اللغات الأصلية ولغات الهدف (المترجم إليها) بمجموعاتهم اللغوية الخاصة وتربط الاختبارات من خلال منظومة المواد التي يمكن إلى حد ما اعتبارها "عامة" لكلتا مجموعتي اللغة. وقد تم اعتبار تطبيقات نظرية إجابة المادة (IRT) لهذا النمط من التصميم واعدة تماماً. كان قد تم استخدام نماذج IRT لربط اختبارات مقررة لمجموعات أحادية اللغة في دراسات متعددة (مثلاً: آنغوف وكوك 1988).

إن النقد الأساسي لدراسات الربط الأحادية اللغة المرتكزة على IRT هو أن تلك الدراسات تصنع افتراضاً غير مستقر حول تكافؤ أجهزة قياس المفردة عند سكان البلدين. بكلمات أخرى، يبدو أن نماذج جهاز قياس المفردة العائدة لـ IRT لم تصمد أمام عينات لغوية مختلفة. وسّع سيرسي (1997) مشكلات استخدام IRT لربط اختبارات لغوية مختلفة. أشار إلى أن "تقديم برهان تجريبي لعدم تغيّر المفردة

عبر اللغات، يتطلب معياراً متفقاً عليه وساري المفعول. إن مقياس الكفاءة (IRT مقياس ثيتا) هو معيار متفق عليه وعرضة للخطأ بسبب عدم وجود مفردات عامة" (صفحة 14). استمر سيرسي في الإشارة إلى أن سير عمليات القياس IRT مثل المعايرة المتزامنة وسير عملية ستوكنغ - لورد (1983) لا يحل المشكلة؛ لأنها تتطلب قياس ما للفروق في الكفاءة بين مجموعتين لغويتين؛ من المحال نظرياً الحصول على هذا القياس دون منظومة مفردات عامة صحيحة.

طريق آخر لتبيان المشكلات المقترنة باستخدام إجراءات IRT للتصاميم أحادية اللغة هو أن تلك الإجراءات تدعي بناء تكافؤ عبر المفردات العامة، وبصورة لا متناهية عبر اختبارات مختلفة مقررة لمجموعات أحادية اللغة.

باستخدام نقاش لين (1993) لتصنيف دراسات الربط، يتم تصنيف النتائج لمعظم دراسات الربط أحادية اللغة، في أحسن الحالات، كدراسات تعديل إحصائية (دراسات تتضمن اختبارات ربط لتراكيب مختلفة) وتكون خاضعة لكل التحذيرات التي يتم تطبيقها حرفياً عند تفسير نتائج دراسات التعديل.

على الرغم من الآراء النقدية لتصميمات مجموعة أحادية اللغة التي أثيرت في وقت مبكر، تجدر الإشارة إلى أن تطبيق نمط التصميم هذا يحدث بصورة متكررة ويقدم غالباً نتائج مفيدة جداً. تدور الموضوعات المقترنة بهذا النمط للتصميم حول تفسير نتائج الدراسة. يتم تفسير تلك النتائج أحياناً كما لو كانت حصيلة دراسة توازن. من المهم أن نلاحظ أنه ببساطة بسبب أن تصميماً متساوياً قد استخدم، لا يعني أن الدراسة قد أنتجت علامات نهائية تعادلية. إن درجات نهائية متساوية، في مفهوم تلك التي تم الحصول عليها من دراسة تعادلية حرفية، تحدث فقط إذا كانت الافتراضات المتضمنة لنموذج التساوي تلتقي مع المعطيات. على أية حال، غالباً ما ينجم عن تطبيقات تصاميم أحادية اللغة درجات نهائية يمكن اعتبارها قابلة للمقارنة بدرجة كافية للأغراض المستخدمة لأجلها.

تصاميم مجموعة ثنائية اللغة:

وصف سيرسي (1997) متغيرات ثلاثة لتصميم مجموعة ثنائية اللغة. الأول هو التصميم الذي تأخذ فيه مجموعة وحيدة ثنائية اللغة للمتقدمين للاختبار كلا الترجمتين اللغويتين للاختبار في ترتيب متساوٍ. أشار سيرسي إلى أنه ربما يكون عائق واحد لهذا النمط من هذا التصميم من آثار التدريب. هذا صحيح على الأخص إذا مثل الاختبارات تكييفات قريبة جداً من اختبار واحد. التصميم ثنائي اللغة الثاني هو التصميم الذي فيه كل واحدة من المجموعات المتساوية ثنائية اللغة تأخذ عشوائياً نسخة من الاختبارات ليتم ربطها. قدم سيرسي الحجة على أن الانطلاق المحتمل لهذا التصميم هو الإمكانية في أنه قد تزول مجموعات عشوائية لأنها ليست متساوية. الثالث هو التصميم الذي تستجيب فيه عشوائياً مجموعات متساوية ثنائية اللغة إلى خليط من مفردات لغة أصلية ولغة هدف.

استمر سيرسي (1997) في القول بأن إحدى المشكلات الرئيسة مع التصاميم ثنائية اللغة هي تعريف "ثنائية اللغة" إجرائياً. من الصعب تعيين متقدمين للاختبار يملكون كفاءة متساوية في كلتا اللغتين موضع الاهتمام، خاصة عندما يعتبر المرء كفاءة اللغة وكأنها تمت إلى التركيب الذي جرى تقييمه. موضوعات إضافية متصلة باستخدام مجموعات ثنائية اللغة في تقييمات لغات مختلفة يجري وصفها بالتفصيل في سيرسي (الفصل 5 من هذا المجلد).

عائق أساسي لتصاميم الربط ثنائية اللغة هو أنه ربما لا تمثل مجموعة ثنائية اللغة أياً من مجموعات أحادية اللغة التي هي المجموعات موضع الاهتمام في الدراسة المقارنة. إن لهذا التحديد مضامين جدية من أجل تعميم نتائج دراسة ربط لغات مختلفة جرى إنجازها باستخدام مجموعة ثنائية اللغة لمجموعات أحادية اللغة.

تصاميم مطابقة لأحادية اللغة:

إنه بإعطاء المشكلات الموصوفة سابقاً مع تصاميم بسيطة لمجموعات أحادية

اللغـة ومـجـمـوعـات ثنائيـة اللغـة، تكون الإمكانيـة في اسـتـخـدام تصـمـيم يطابق مجـمـوعـات منفصلة أحـادية اللغـة على بعض المتـغـيـرات التي ربما تؤثر على نتـائج الربط تكون الإمكانيـة مغـرية تماماً. على أية حـال، تصـاميم كهـذه نادراً مـا جـرى استخدامها بنجاح. تحاول التصاميم المطابقة لأحادية اللغة أن تتجاوز الحاجة إلى مواد عامة لكي يقيم الفرق في المهارات/ القدرات وذلك باستخدام مجموعات لأجل دراسـة الربط الذي يتفق مع الآراء النقـدية الوثيـقـة الصلة أياً كـانت المهـارات أو القدرات التي يجري تقييمها بواسطة اختبارات لغة مختلفة.

كمـا أشـار سيـرسي (1997)، كـان قد جـرى التحـري بشكل مـوسع تمامـاً عن تأثيـرات مـجـمـوعـات مطابقـة في تصـاميم مـتـعـادلة لأشكال تقليدية (انظر كوك، إيغنور، وشـمـيت 1989، إيغنـور، سـتـوكنغ، وكـوك 1990، كـولن 1990؛ ليـفنغـسـتـون، دورانز، ورايت 1990؛ سكاغز 1990). لقد جرى خلط نتائج هذه الدراسات. اقترح ليفنغستون وآل أنه ربما يتم تحسين التساوي عبر التطابق في ميل نزوع الدرجات النهـائيـة (روزبنـون وروبن 1983)، بينمـا حـذّر كـوك وآل من تقنيـات كـهـذه. إن التحذيرات نفسها المذكورة عند تقييم استخدام مجموعات ثنائية اللغة من أجل دراسات ربط لغات مختلفة ينبغي أن تذكر في سياق استخدام مجموعات متطابقة لأجل أنماط دراسات الربط تلك.

نقطة رئيسة أثارها لين (1993)، عند مناقشة دراسة الربط للتصنيفات، وكانت في أن كل دراسات ربط أخرى غير دراسات التساوي الصحيحة تعاني من مشكلة اعتماد النتائج على المجموعة. بناء على ذلك لا يمكن تعميم النتائج لدراسة ربط لغات مختلفة تم إنجازها باستخدام مجموعات أحادية اللغة مبنية بحيث إنها تتطابق مع متغيرات خاصة هي مفتاح القدرة المقاسة. لا يمكن تعميم تلك النتائج على مجموعات أحادية اللغة متغايرة الخواص بدرجة أكبر والتي تكون بشكل نهائي المجموعات مركز الاهتمام.

في القسم التالي من هذا الفصل تجري مناقشة دراسات ربط ثلاث جرى العمل بها عبر العشرين سنة الماضية بغرض ربط الدرجات النهائية في SAT و PAA يتم نقد كل دراسة من منظور المناقشة السابقة حول تصاميم الربط.

تطوير علاقة بين الدرجات النهائية لـ PAA** وSAT*

لقد تم تصميم سلسلة دراسات جرت لتطوير علاقة بين الدرجات النهائية لاختبار SAT والدرجات النهائية لاختبار PAA لتطوير مقياس عام يسهّل مقارنات الدرجات النهائية المحصّلة في الاختبارين.

كان الباحثون العاملون على كل الدراسات مدركين أن الفروق الأساسية في اللغة، والعادات، والقيم ربما تضعف بصورة ممكنة مقارنات بين مجموعات تأخذ الاختبارين. على أية حال، كان أولئك الباحثون ملتزمين بتطوير علم منهج مثالي يمكن استخدامه لبناء مقياس غير منحاز على قدر الإمكان.

من المهم في تلك النقطة التأكيد على أن اختبارات PAA ليست ترجمة مباشرة أو تكييف لـ SAT بالرغم من أن اختبارات PAA مصممة لقياس التراكيب نفسها مثل الـ SAT، فاختبارات PAA تحتوي على مفردات مختلفة ويتم تطويرها بشكل مستقل تماماً عن اختبارات SAT تم اتخاذ قرار من قبل مجلس الكلية، باكراً في تاريخ برنامج الاختبار لـ PAA، أنه بسبب التعقيدت والصعوبات المتضمنة في تكييف اختبار من إحدى اللغات إلى الأخرى، يكون من الأفضل حفظ "التساوي" بين الاختبارين إذا كان كل اختبار مصمماً لقياس "التركيب نفسه" لكن بلغة مختلفة.

هناك ظاهرة مميزة لاختبارات PAA وهي أنه جرى تصميمه ليستخدم في مضامين متعددة للناطقين بالإسبانية. فسكان بلاد «الهسيبانك» المتنوعين، على

(*) Scholastic Assessment Test (SAT).

(**) Prueba de Aptitude Academica (PAA) (AST النسخة الإسبانية للاختبار الأمريكي)

سبيل المثال، المكسيكيون البروتريكيون يختلفون الواحد منهم عن الآخر بدرجة كبيرة بالطريقة نفسها، مثلاً لنقل اختلاف رعايا الولايات المتحدة ورعايا بريطانيا العظمى. تتكلم كلتا هاتين المجموعتين الإنكليزية، لكن الفوارق الطفيفة في اللغة تختلف في بلدان مختلفة. جرى نقل تحليلات متفاوتة لتوظيف المفردة (DIF) إلى الـ PAA لتأكيد صدق البنية التي يقيسها الاختبار عبر سكان بلاد الهسبانك المختلفين (انظر مثال، سيرسي وآلالوف 2003).

كانت الدراسة الأولى التي أجراها آنغوف ومودو (1973) في خريف 1971 قد تم توجيهها من أجل غرض ربط الدرجات النهائية في الـ PAA بالدرجات النهائية في الـ SAT وقد جرى استخدام نتائج دراسة آنغوف/ مودو لمقارنة الدرجات النهائية في الـ PAA بالدرجات النهائية لـ SAT لحوالي مدة عشر سنوات. قادت التقدمات في التقنية، كما في التحقق من أن تدريباً جيداً لتكرار ومراجعة نتائج دراسات الربط بشكل دوري. قادت إلى تكرار ربط SAT/PAA بدراسة آنغوف وكوك (1988). اتبعت دراسة آنغوف - كوك التصميم الأساسي للدراسة الأبكر، ولكنها استبدلت نظرية الاختبار الكلاسيكية لعلم المنهج بتقنيات IRT جرت إدارة دراسة الربط الأكثر حداثة والمعتمدة من الـ PAA والـ SAT من قبل شميت، دورانز، ماغرينا، وكوك (1998).

وكان الغرض من هذه الدراسة تقديم عامل ربط حديث للاختبارين الذي عكس تغييرات حديثة في مواصفات خواص الاختبار. وظفت الدراسة الثالثة علم منهج مختلف تماماً من أجل ربط الاختبارين من علم المنهج الذي استخدم في الدراستين السابقتين. ما يتبع هو نقاش موجز ونقد للدراسات الثلاث للربط.

دراسة آنغوف-مودو:

طوّر آنغوف ومودو (1973) علم منهج لتأمين تحويل درجات نهائية رياضية ولفظية باللغة الإسبانية PAA إلى درجات نهائية رياضية ولفظية خاصة بالـ SAT جرى تقرير كلا الاختبارين على طلاب مدرسة ثانوية من أجل أغراض القبول في

كلية. وكما جرى ذكره سابقاً، بالرغم من أن PAA و SAT تشاركت في البنية نفسها والشكل نفسه، فقد تم تأليف كل منهما من مفردات أصلية مستقلة، مما يعني أن الاختبارات لم تكن تراجم مكيّفة الواحدة منهم من الأخرى. كان الغرض من الدراسة المنفذة من قبل آنغوف ومودو تأمين لوائح تحويل بين الـ PAA والـ SAT تسهل مقارنات مباشرة للمجموعات الفرعية من مجموعتي اللغة اللتين أخذتا اختباراً ملائماً بلغاتهم الأصلية. بالإضافة إلى ذلك، كان من المتوقع أن تساعد لوائح التحويل في تقييم الأرجحية للنجاح في كليات البلد الرئيس الذي ربما تم الحصول عليه من قبل طلاب من بورتوريكو.

تألفت دراسة آنغوف ومودو (1973) من جانبين. تضمن الجانب الأول انتقاء مفردات "عامة" تمّ استخدامها كاختبار معتمد في دراسة "التوازن" ويتألف الجانب الثاني من "تساوٍ" فعلي. كان على الطريقة المستخدمة في الجانب الأول أن تختار منظومتي مفردات، الواحدة بالإنكليزية أصلياً والثانية بالإسبانية أصلياً، وأن تترجم كل منظومة إلى اللغة الأخرى. وبعد الترجمة، تم تقرير منظومتي المفردات (واحدة بالإسبانية وواحدة بالإنكليزية) على طلاب ملائمين أحادي اللغة من أجل أغراض اختبار تمهيدي. تم تنفيذ إدارات الاختبار التمهيدي لهذه الدراسة في خريف عام 1970، وعلى أساس تحليل معطيات الاختبار التمهيدي، جرى انتقاء منظومتي مفردات، الواحدة لفظية والثانية رياضية، كمنظومات مفردات "عامة" لكي يجري استخدامها من أجل "عمليات تساوٍ" رياضية ولفظية خاصة.

في "التساوي" من الدراسة، جرى تقرير المفردات "العامة"، الظاهرة في كل من الإسبانية والإنكليزية، بلغة ملائمة إلى جانب ومع الشكل الإجرائي لـ PAA في نوفمبر 1971، ومع الإجرائي لـSAT في يناير الثاني 1972 وقد تم استخدام المعطيات من تلك الإدارات لتوجيه كل عمليات التساوي الخطية و Equipercentile ذات التصنيف المئوي لقيم المتغير إلى الاختبارات الرياضية واللفظية لـ PAA وSAT.

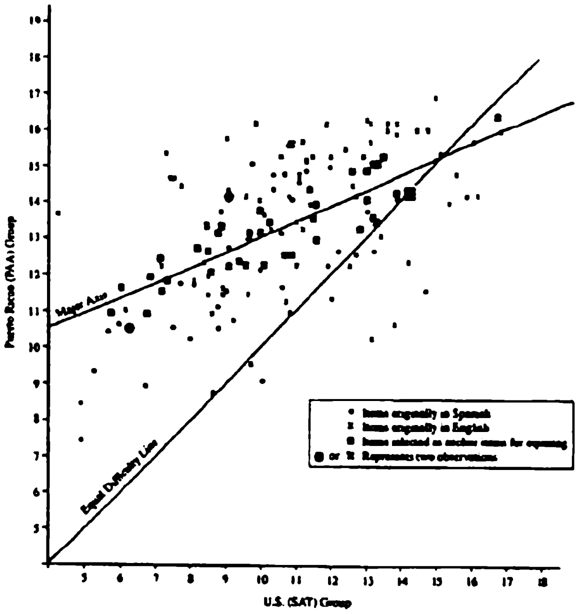
هناك مظاهر عدة من هذه الدراسة المبكرة تستحق الوصف بالتفصيل. تتألف المرحلة I للدراسة من بناء مفردات "عامة" أو اختبار رابط يستخدم لتقييم الفروق في المقدرة بين المجموعة PAA الناطقة بالإسبانية والمجموعة SAT الناطقة بالإنكليزية. وقد تم رسم إطار المجموعة الأولية للمفردات التي سيجري استخدامها لتشكيل اختبارات الربط في أعداد متساوية تقريباً من أطر مفردات الـ PAA والـ SAT. جرت ترجمة تلك المفردات إلى اللغة الثانية من قبل مجموعة صغيرة من خبراء ثنائي اللغة. جرى بذل جهد لإنتاج منظومة مفردات، بالإنكليزية والإسبانية، التي كانت، تقريباً على قدر الإمكان، مساوية بالمعنى في اللغتين. وفي وقت لاحق جرت إعادة ترجمة كل المفردات إلى لغتهم الأصلية ومقارنة الترجمات المعادة (النسخ التي خضعت إلى ترجمتين) مع النص الأصلي.

ومن ثم جرى اختبار تمهيدي لمجموعة المفردات العامة وذلك بتطبيقها على مجموعات طلاب تقدموا لكل من اختباري الـ PAA أو الـ SAT باللغة المناسبة. وتلي الاختبار التمهيدي، عملية غربلة المفردات إحصائياً وذلك برسم العلاقة البيانية بين صعوبة المفردات وقيم دلتا لكل من المفردات الرياضية واللفظية المأخوذة من قبل مجموعات ناطقة بالإنكليزية وبالإسبانية. (انظر آنغوف ومودو، 1973، من أجل وصف "المثلث"). كان الغرض من رسوم دلتا البيانية هو التمكّن من تعيين المفردات التي تملك معنى مختلف للمجموعتين، PAA و SAT وقد اعتبرت المفردات ملائمة بدرجة متساوية للمجموعات الناطقة بالإنكليزية كما هي للناطقة بالإسبانية على أساس قربهم من المحور الرئيس للقطع الناقص من رسم دلتا البياني كما في الأشكال 6-1 و 6-2 المأخوذة من آغنوف ومودو (1973)، توضح نتائج الرسوم البيانية المثلثية لمنظومات مفردات الربط الرياضية واللفظية.

نقطة مهمة ينبغي ملاحظتها وهي أنه عند مقارنة الرسم البياني للمفردات اللفظية مع الرسم البياني للمفردات الرياضية نجد درجة أعظم من التشتت في المفردات اللفظية عن المحور الأساسي في القطع الناقص لرسم دلتا البياني.

فسّر آنغوف ومودو (1973) التشتت الكبير للمفردات اللفظية كمشير على أن المفردات اللفظية لا تملك تماماً المعنى النفسي لمجموعتي اللغة. وتابعا القول في أن التشتت كان كافياً لإبقاء الشك حول نوعية أي توازن تم القيام به مع تلك المفردات. جرى تحسين الوضع بحذف المفردات الأكثر شذوذاً؛ على أية حال، استمر المؤلفان في إبداء الاهتمام بأن تفاعلات المجموعة بالمفردات والمشار إليه في المعطيات الموجودة في رسوم دلتا البيانية قد جعل التوازن "أقل جدارة كثيراً بالثقة مما هو متوقع من تعادل اختبارين متساويين جرى إعدادهما من أجل أعضاء في ثقافة اللغة نفسها" (صفحة 14).

الوجه II من الدراسة يتألف من "تساوٍ" فعلي لـ PAA و SAT، باستخدام منظومة مفردات الربط المبنية في الوجه I جرى انتقاء أربعين من المفردات اللفظية وخمس وعشرين من المفردات الرياضية كمفردات عامة لأجل الوجه II من الدراسة. تمّ إقرار المفردات، في موازاة ومع نسخة معدلة إجرائية للاختبارات الخاصة في نوفمبر 1971 الإدارة لـ PAA، ويناير 1972 الإدارة لـ SAT إجراءات «تساو» متفقة مع القواعد المقررة التي جرى استخدامها لربط الدرجات النهائية في الـ PAA مع الدرجات النهائية في الـ SAT وقد جرى استخدام مقياس النشاط الطولي الخطي (Tucker) .

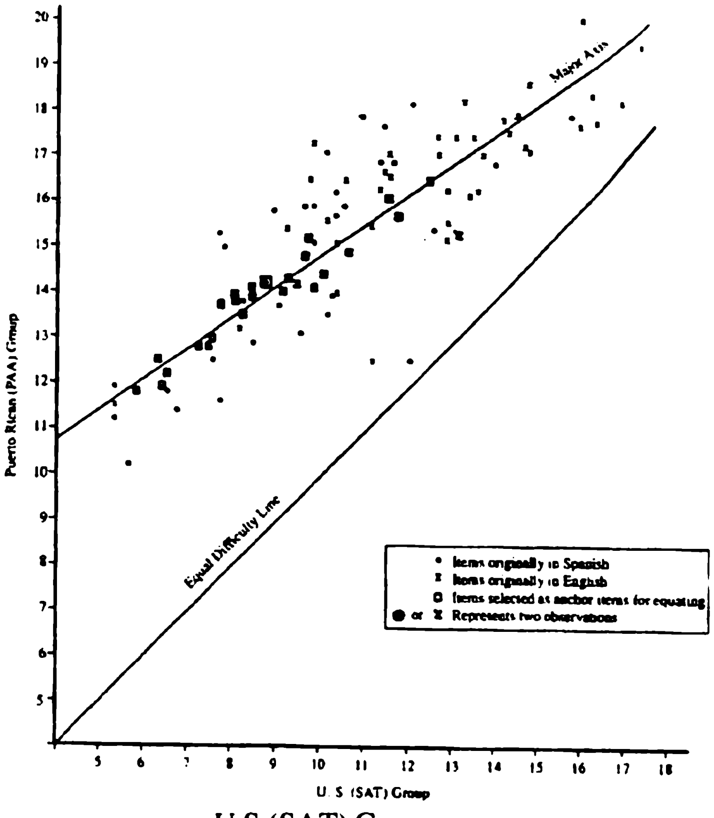


U.S (SAT) Group

الشكل 1.6 رسم دلتا البياني للمفردات اللفظية.

ومقياس التصنيف المئوي لقيم المتغير (آنغوف، 1984؛ تجري الآن الإشارة إليه كمقياس تصنيف مئوي "متسلسل")، وسير العملية الطولي لـ ليفن (1955). نظراً لأن المعطيات لا تقابل أياً من نماذج التوازن الثلاث المتعارف عليها، تقرر إيجاد معدل نتائج النماذج الثلاث، بإعطاء وزن أعظم لنتائج التصنيف المئوي لقيم المتغير. تظهر الأشكال 306 و 406 رسوماً بيانية متبعثرة للتعادلات الرياضية واللفظية.

أشارت نتائج "التساوي" اللفظي إلى أن مقياساً وسطياً قيمة (500) لـPAA كان مساوياً لمقياس لفظي درجة نهائية (350) لـ SAT فعلياً أدنى من القيمة الوسطية. أشارت النتائج في "تساوي" الرياضيات إلى أن الدرجة النهائية لـPAA لـ 500 نتجت حتى في أخفض درجة رياضيات نهائية لـ (319) SAT.



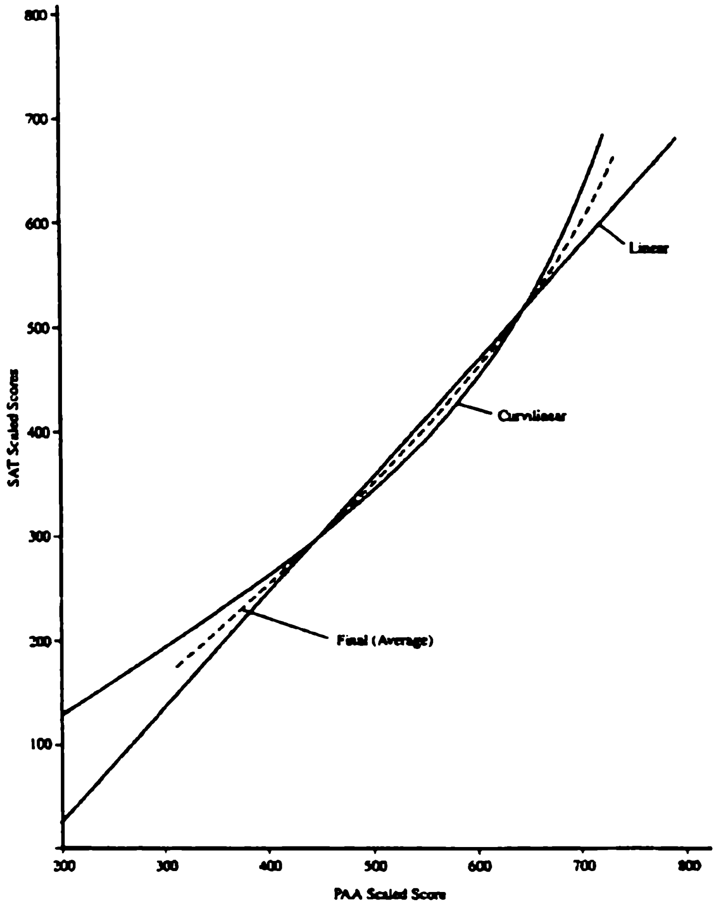
U.S (SAT) Group

الشكل 2.6 رسم دلتا البياني لمفردات الرياضيات.

حذّر آنغوف ومودو (1973) من أن "الدقة في تلك التحويلات هي محددة بملاءمة الطريقة المستخدمة في اشتقاقها والمعطيات المجمعة في أثناء مجرى الدراسة، من المأمول أن تكون تلك التحويلات مفيدة في تنوع للمحتويات لكن... لكي تكون مفيدة ستحتاج في كل لحظة إلى أن تعزز بمعطيات إضافية متميزة للمحتويات" (صفحة 41).

دراسة آغنوف-كوك:

استخدمت الدراسة التي قام بها آنغوف وكوك (1988) التصميم الأساسي نفسه مثل ذلك الذي جرى استخدامه من قبل آنغوف ومودو (1973).

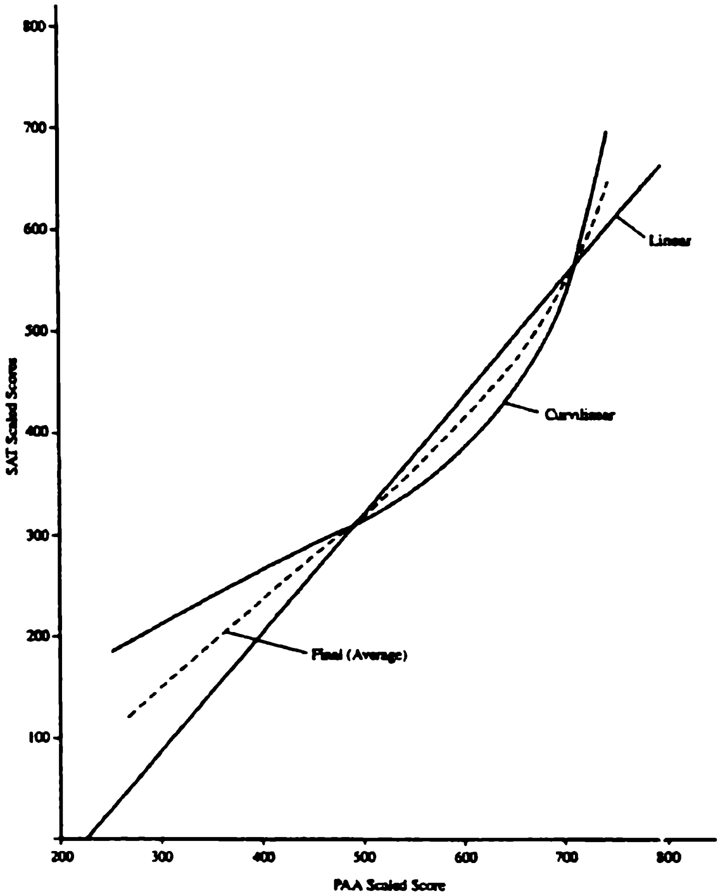


الدرجات النهائية المقاسة PAA

الشكل 6-3 نتائج التعادل للاختبارات اللفظية.

لكنها وظّفت علم المنهج IRT لتأخذ مكان كل من تشتت مفردة الرسم البياني المثلثي المستخدمة لانتقاء اختبار المفردات العامة ومن علم منهج التساوي المتعارف عليه والمستخدم في الدراسة الأبكر.

على وجه شبه الدراسة السابقة، جرى القيام بدراسة آنغوف وكوك (1988) في جانبين. تألّف الجانب الأول من انتقاء



الدرجات النهائية المقاسة PAA

الشكل 6 - 4 نتائج التعادل لاختبارات الرياضيات

مفردات الربط التي سيجري استخدامها في الجانب II، وجه "التساوي" للدراسة. لأجل دراسة آنغوف وكوك، تمّ اتباع علم المنهج المؤسس في دراسة أسبق من أجل التكييف، إعادة التكييف، الاختبار التمهيدي للمفردات. كان الفرق بين

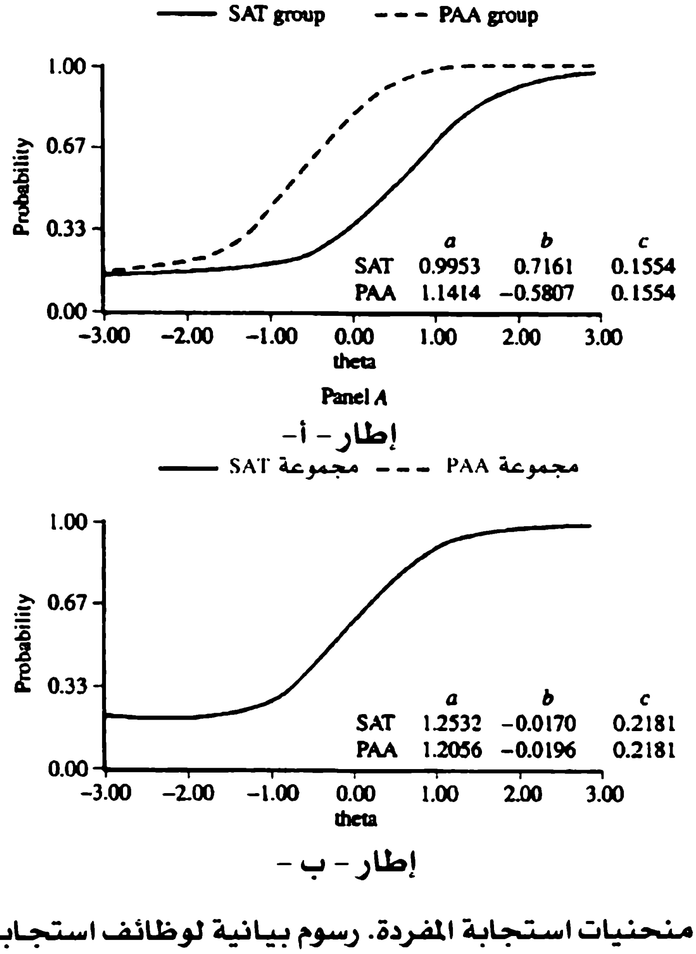
الوجه I في الدراستين هو علم المنهج المستخدم في تشتت المفردات لأجل الانتقاء كمفردات "عامة". قيّم آغنوف وكوك المفردات بمقارنة الفروق، بصرياً وإحصائياً، بين المنحنيات الخاصة للمفردة IRT (ICCs).

يعطي الشكل 6-5 أمثلة على الرسوم البيانية لـ ICCs من أجل مفردات رياضيات ولفظية جرى تقديرها لمجموعات ناطقة بالإنكليزية، وبالإسبانية. يظهر الإطار (أ) في الشكل 6-5 أنه من أجل كل مستويات القدرة (Theta) تملك مجموعة PAA احتمالاً أعلى في الحصول على إجابة صحيحة للمفردة منه في احتمال مجموعةSAT مفردة كهذه لا يمكن بوضوح اعتبارها مفردة "عامة" للمجموعتين وبالتالي سقطت أثناء جانب تبعثر المفردة في الدراسة. يحتوي الإطار "ب" في الشكل 6-5 مقارنة لـ ICCs المحصلة لأجل مادة رياضيات معطاة إلى مجموعات SAT و PAA.

بالمقابلة مع المنحنيات الظاهرة على الإطار (أ)، تكون ICCs لأجل مفردة رياضيات معطاة إلى مجموعتين من المتقدمين للاختبار تقريباً مطابقة؛ وذلك يعني أن الأفراد بكافة مستويات القدرة في كلتا المجموعتين يملكون الاحتمال نفسه في الحصول على إجابة صحيحة للمفردة. ولا تفضّل المفردة أياً من المجموعتين. مفردات كهذه المفردة يمكن أن تعتبر مثالية من أجل الإدخال في منظومة المفردات "العامة" المستخدمة في ربط اختباري الرياضيات.

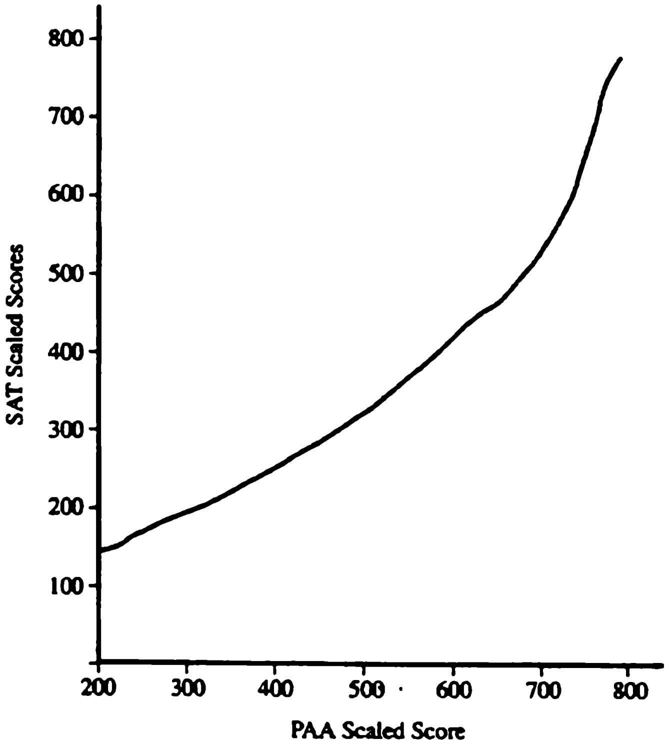
يحاذي جانب "التساوي" في الدراسة نظيره في دراسة آنغوف ومودو (1988) باستثناء استخدام إجراءات IRT تم إقرار منظومات مفردة "عامة" لفظية ورياضيات إلى الأمام مع اختباراتهم الإجرائية الخاصة (SAT) أو (PAA) على مجموعات ملائمة أحادية اللغة. جرى جمع معطيات SAT في كانون الأول 1985 ومعطيات PAA في الإدارة في تشرين الأول 1986 كانت طريقة التساوي الـ IRT المستخدمة في هذه الدراسة هي التساوي لـ IRT المطبق (كوك وإيغنور، 1983؛ بيترسن، كوك، وستوكنغ، 1983) فقط تم الإبلاغ عن نتائج التساوي المشكلة بخط منحن لـ IRT من أجل الدراسة. يجري تمثيل هذه النتائج لاختبارات الرياضيات واللفظية في الأشكال 6-6 و6-7 .

من الواضح من مراجعة الخطوط البيانية المعروضة في الأشكال 6-6، 6-7، أن العلاقة بين مقاييس الـ PAA والـ SAT متسمة بخطوط منحنية بشكل ملحوظ. هذه كانت أيضاً الحالة بالنسبة إلى نتائج تعادل محصّلة في دراسة آنغوف ومودو (1973)؛ على أية حال، اختار آغنوف ومودو أن يوجدا معدل النتائج المشكّلة خطاً منحنياً مع النتائج الخطية. وتشير نتائج دراسة آنغوف وكوك (1988) إلى أن الفروق بين مقاييس الـ PAA والـ SAT حسب الدرجة النهائية 500 لـ PAA كانت حوالي 180 إلى 185 علامة. أشارت نتائجهما إلى فروق مشابهة بالنسبة لربط الرياضيات، وهذا يعني، في الدرجة النهائية PAA من 500، كانت الفروق بين مقياس الـ PAA والـ SAT حوالي 180 إلى 185 علامة.



الشكل 6-5 منحنيات استجابة المفردة. رسوم بيانية لوظائف استجابة المفردة من أجل المفردات اللفظية (الإطار أ) والرياضية (الإطار ب) مفردات معطاة إلى مجموعات SAT ومجموعات PAA، موضحة اتفاق جيد أو ضعيف بين المجموعات.

سلمت نتائج الدراسة لـ آنغوف وكوك (1988) تحويلات الدرجات النهائية اللفظية PAA إلى المقياس اللفظي SAT أدنى جوهرياً من الدراسة الأبكر، على الأخص في المجال الأوسط لمقياس الدرجة النهائية. أظهرت التحويلات إلى المقياس الرياضي لـ SAT اتفاقاً أفضل مع النتائج الأبكر. فكّر المؤلفون أن الفروق في النتائج قد تعزى إلى فروق في علم المنهج أو إلى صعوبات متأصّلة لاختبارات "التساوي" المثقلة لفظياً نسبة إلى مجموعات لغة مختلفة.



شكل ٦،٦ نتائج تعادل الاختبارات اللفظية

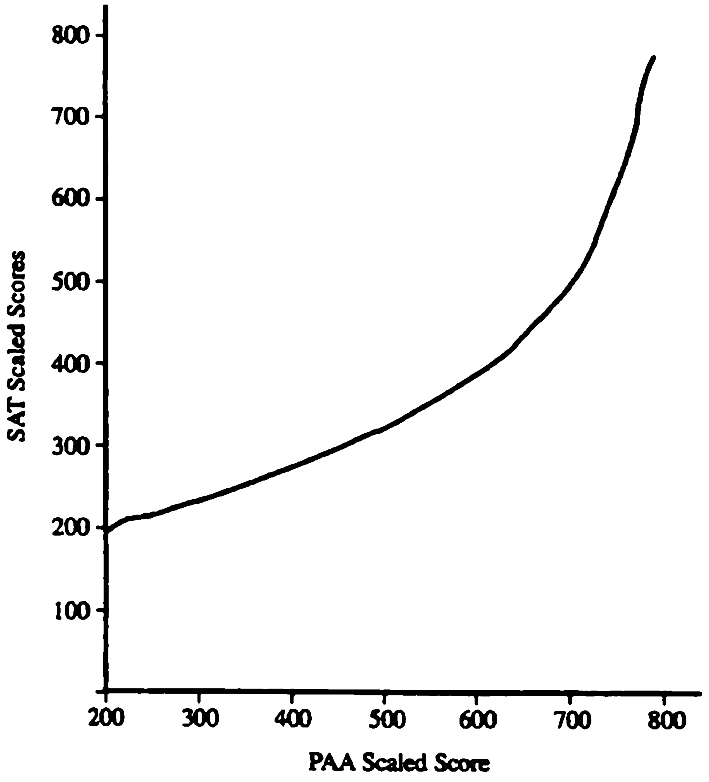
كتابة نقدية للدراستين السابقتين:

إن علوم المنهج لربط الدرجة النهائية الموظّفة في الدراستين، سير عمليات التساوي المتّسمة بخطوط منحنية وسير عمليات توكر وليفين والتصنيف المئوي لقيم المتغير، هي سير عمليات معقول للاستخدام إذا كان السؤال موضع الاهتمام هو مقارنة توزيعات الدرجات النهائية لمجموعتين من المتقدمين للاختبار (أو في حالة

طرائق خطية، الحركتين الأوليتين للتوزيع). تذكر أن الأغراض الأصلية لربط /SAT PAA، كما جرى وصفها من قبل آنغوف ومودو (1973)، كانت:

(أ) أن نقارن توزيعات الدرجات النهائية للمجموعات الفرعية من سكان بلدين.

(ب) أن نقيّم النجاح المحتمل لطلاب بورتو ريكان الذين كانوا مؤخراً مهتمين بالمواظبة على الحضور في كليات البلد الرئيس. وكانوا يخضعون إلى الدرجات النهائية لـ PAA من أجل أغراض القبول.



الشكل 6 -7 نتائج تساوي اختبارات رياضيات

بالاعتماد على مدى الجودة التي لاقت فيها المعطيات افتراضات النماذج الإحصائية المستخدمة لأجل الربط، ومدى الجودة التي تمّ بها تنفيذ الربط، من الممكن أن علم المنهج الموظّف في دراسات آنغوف ومودو (1973)، وآنغوف وكوك

(1988)، يستطيع أن يقدم أساساً معقولاً لمقارنات وتوزيعات الدرجات النهائية. سبب ذلك، أن الغرض من إجراءات IRT، الخطية والتصنيف المئوي لقيم المتغير، الموظّفة في هاتين الدراستين هو تحويل توزيع الدرجات النهائية المحصّلة في اختبار واحد لتلائم تلك المحصّلة في الاختبار الثاني، لأجل مجموعة خاصة من المتقدمين للاختبار.

بالرغم من أنه يمكن نظرياً لتقنيات إحصائية مستخدمة في هذه الدراسات أن تقدم نتائج نهائية تقابل هدف الدرجات النهائية المقارنة، من المحتمل ألا يقابل الهدف من قبل نتائج الدراستين بسبب طبيعة المعطيات المستخدمة لأجل الربط. على أية حال، سيقدم الإجراء المستخدم حلاً للمشكلة شريطة أن تتلاقى الافتراضات الأساسية، في علم المنهج.

خذ بعين الاعتبار السبب الثاني لتنفيذ الربط SAT/PAA، وهذا يعني تقييم نجاح الطلاب من بورتوريكو في كليات البلد الرئيس والجامعات، كيف يكون بالإمكان استخدام العلاقة المؤسسة بين الـ PAA و SAT في الدراستين السابقتين للتنبؤ بمدى الجودة التي ربما يحققها طالب ثانوي في سان جوان عندما ينتظم هو أو هي في كلية، لنقل، في ميامي! ماذا يعني أن نقول إن طالباً حصل على 500 في الـ PAA سيحصل على درجة نهائية 320 في الـ SAT إذا لم يتكلم الطالب الثانوي في سان جوان الإنكليزية، بالتأكيد لن يحصل هو أو هي على درجة نهائية 320 في الـSAT وسيقضي وقت دراسة شديد الصعوبة في كلية أو جامعة في الولايات المتحدة.

بحثت بينوك - رومان (1995) العلاقة بين الدرجات النهائية لاختبارات قبول مستوى الخريجين معطاة بالإنكليزية وبالإسبانية لمجموعة من الطلاب الذين كانوا أكثر تمكناً بالإسبانية منهم بالإنكليزية ووصلت إلى أن التمكن من الإنكليزية ساهم في درجات الطالب النهائية في اختبار سجل الخريجين (GRE) اللفظي واختبارات علم الأحياء، علم النفس، التحليلية، الرياضيات. وجدت بينوك-رومان أن التمكن من الإنكليزية ساهم بشكل مختلف معتمداً على مستوى تمكن الطالب ومعتمداً على

فحوى الاختبار. وقد ناقشت حقيقة أنه من الممكن لطالب موهوب في اللغة الثانية أن يحصل على درجة نهائية دون المعدل في الاختبار اللفظي GRE ببساطة بسبب فهم قراءة أبطأ.

إن المعاني المتضمنة في دراسة بينوك - رومان (1995) من أجل دراسات ربط (مقارنة) SAT/PAA هي أنه إذا كان الغرض الرئيس من الدراسة هو تقديم وسيلة من أجل تقييم مدى الجودة التي سيؤديها طالب في بورتوريكو في كلية البلد الرئيس أو جامعة، ثم قياس ما للمقدرة باللغة الإنكليزية يجب أن يؤخذ بالاعتبار. يحاول علم المنهج المستخدم لأجل دراسة الربط SAT/PAA الموصوف لاحقاً أن يأخذ هذه المعاني المتضمنة في الحسبان.

دراسة شميت، دورانز، ماغرينا، وكوك:

في ربيع 1994، جرى إدخال SATI جديد. يحتوي الاختبار الجديد على أنماط مفردات جديدة وتم بناؤه وفقاً لمواصفات إحصائية ومحتوى منقّح. (انظر كوك، 1995)، لأجل وصف مراجعات لـ (SAT) في أكتوبر 1996، تم أيضاً تنقيح الـ PAA لتشمل أنماط مفردات جديدة، محتوى منقّح، ومواصفات إحصائية. تغييرات في الـ PAAموازية للتغيرات المدخلة في SATI على الأخص، لم تشمل الـ PAA اللفظية الجديدة على مفردات مناقضة وتملك نسبة مئوية أعلى من المفردات اللفظية التي تتصل بمقاطع قراءة نقدية (56 ٪ مقابل 31 ٪). بالإضافة إلى المفردات المتعددة الاختيار التقليدية، يتضمن الـ PAA الجديد للرياضيات مفردات حيث ينتج المتقدم للاختبار استجابته أو استجابتها الخاصة. (انظر مجلس الكلية، 1995، لوصف موسّع للتغييرات في PAA الجديد). الفرق الوحيد بين الـ SATI والـ PAA الجديد هو أن الـ SATI يسمح باستخدام الآلات الحاسبة في اختبار الرياضيات.

قدمت دراستا الربط SAT/PAA السابقتين جداول اتفاق بين الدرجات النهائية في نسخ سابقة لـ PAA و SAT آنغوف وكوك، 1988؛ آنغوف ومودو،

1973 استخدمت دراسات القياس هذه مفردات "عامة" لتتلاءم مع أية فروق بين مجموعات الـ SAT و PAA. وبسبب الموضوعات الموصوفة باكراً، اقتربت آخر دراسة من التشابه في الدرجات النهائية بين الـ SATI والـ PAA من منظور مختلف. تم الحصول على جداول اتفاق بالرغم من أن طرائق "التساوي" المستخدمة في الدراستين الأبكر تدّعي أن الاختبارين كانا جوهرياً أشكال بديلة تمثّل البناء نفسه. نظراً لكون الـ PAA باللغة الإسبانية والـ SATI باللغة الإنكليزية وتحتوي كل منهما على مفردات تمّ تطويرها بصورة دقيقة واختبارها مسبقاً على السكان المعنيين الخاصين في كل بلد، فليس بالإمكان اعتبار الـ PAA والـ SATI أشكالاً بديلة.

لأجل الدراسة الثالثة PAA/SAT من قبل شميت وآل (1998)، جرى استخدام طريقة تنبئية. جرى افتراض أن تطوير جدول الاتفاق لم يكن أساسياً لنجاح الدراسة. بناء على ذلك قدمت طريقة التنبؤ الموظّفة قياس مقدرة اللغة الإنكليزية، اختبار الإنجاز للإنكليزية كلغة ثانية (ESLAT) ، التصميم الأساسي لأجل الدراسة.

دراسات سابقة وثيقة الصلة بهذا البحث، تم استخدام طريقة الارتداد من قبل ألدرمان (1981) وبولدت (1969) لدراسة العلاقة بين اختبارات معطاة بالإنكليزية وبالإسبانية إلى مجموعات ثقافية مختلفة. في دراسة ألدرمان، جرى اختبار طلاب على الـ PAA، (SAT TOEFL اختبار الإنكليزية كلغة ثانية)، و ESLAT جرى اعتبار التمكن من اللغة في (TOEFL) أو (ESLAT) كمتغير وسيط في التنبؤ بنتائج اختبار SAT من نتائج PAA نجم عن التمكن الأعلى من اللغة، كما تم قياسه بواسطة تلك الاختبارات، علاقة أقوى بين الدرجات النهائية لـ PAA و SAT تحدد تلك النتائج أهمية استخدام قياس التمكن من اللغة عند خلق معادلة تنبؤ بين SAT و PAA نظراً لأن جميع المتقدمين لاختبار PAA يأخذون أيضاً اختبار إنجاز الإنكليزية كلغة ثانية (ESLAT) من أجل أغراض القبول، يمكن أن تستخدم نتائج ESLAT كمتغير وسيط للغة. كانت عينة الدراسة الثالثة PAA/SAT شميت وآل، 1998 مرشحين متوفرين أخذوا اختبار PAA الجديد في بورتوريكو من يناير

الثاني 1996 إلى يونيو 1997، في التحاليل تم فقط اعتبار الدرجة النهائية الأخيرة للطلاب الذين أعادوا الاختبار ضمن مدة زمنية محددة. كان لكل طالب مشمول بالدراسة درجات الاختبار النهائية التالية:

(أ) SATI اللفظية والرياضيات، من يناير 1996 على يونيو 1997 (تقديم الاختبار).

(ب) PAA اللفظية والرياضيات، من أكتوبر 1996 إلى يونيو 1997 (تقديم الاختبارات).

(ج) ESLAT، من أكتوبر 1996 إلى يونيو 1997 (تقديم الاختبارات).

جرى اعتبار كل من نماذج التنبؤ* الخطي والمنحني من أجل التنبؤ بالدرجات النهائية SATI من الدرجات النهائية PAA وESLAT.

وجد شميت وآل (1998) أن الارتباطات بين SATI و ESLAT/PAA في العينة تستحق الملاحظة بدرجة كبيرة. بلغ معامل الارتباط بين PAA رياضيات وSATI رياضيات 0.82 مشيراً إلى أن الاختبارات تقيس تراكيب بنيوية متشابهة، لكن ليست نفسها. إضافة إلى ذلك، ارتبطت SATI رياضيات بـ 0.57 درجة مع ESLAT، ارتباط يشير إلى أن ESLAT تعمل كمتغير وسيط للتمكن من اللغة من جل الدرجات النهائية للرياضيات. الجدول 6-1 يحوي الارتباطات لدرجات الاختبار النهائية تلك.

Test Score	*ESLAT*	*PAA-M*	*PAA-V*	*SAT-M*	*SAT-V*
ESLAT	1.00	.51	.45	.57	.74
PAA-MATH	.51	1.00	.61	.82	.60
PAA-VERBAL	.45	.61	1.00	.56	.62
SAT-MATH	.57	.82	.56	1.00	.69
SAT-VERBAL	.73	.60	.62	.69	1.00

جدول 6-1 الارتباطات بين SAT و PAA اللفظي والرياضيات وال ESLAT

(*) المقصود استعمال:
Linear and Curvilinear multiple regression models.

تظهر الارتباطات بين ESLAT و SATI اللفظي اقتراحاً أقوى لمتغير وسيط للتمكن من اللغة، بالنسبة إلى SATI الارتباط 0.73 درجة مع ESLAT وأقل بدرجة معتبرة، 0.62 مع PAA اللفظي. لاحظ أن PAA رياضيات تملك فقط ارتباطاً أدنى قليلاً مع PAA اللفظي (0.60) من الارتباط الذي أظهره SATI اللفظي وPAA اللفظي. من الواضح من هذه المعطيات أن أي جدول ارتباط جرى تطويره مستخدماً اختبارات بهذا الترتيب لدرجات الارتباط ستكون له قيمة مثيرة للتساؤل، بغض النظر عن علم المنهج المختار للربط. كانت إحدى الأسئلة لهذه الدراسة فيما إذا استطاع PAA اللفظي أم لم يستطع إضافة الكثير إلى تنبؤ الدرجات النهائية اللفظية SATI إلى درجة أبعد من الذي استطاعت ESLAT فعله بنفسها.

جرت مقاربة معادلات للتنبؤ بدرجات نهائية SAT رياضيات و SAT لفظية من درجات نهائية PAA ودرجات نهائية ESLAT بالصيغة التالية لـ SAT اللفظي:

SAT اللفظي المقدّر = 371. * PAA اللفظي + 797. * ESLAT-284

لاحظ أن الوزن المعين لـ ESLAT هو أكثر مرتين من ذلك المعين لـ PAA اللفظي. المعادلة التقريبية للتنبؤ بـ SAT رياضيات من PAA رياضيات وESLAT تجعل من الواضح أن PAA رياضيات هو عامل التنبؤ الأكثر أهمية:

SAT رياضيات المقدر= 688. * PAA رياضيات + 259. * ESLAT - 150

من المستطاع رؤية أن PAA رياضيات تملك وزناً أكثر من مرتين بالسعة من ذلك الذي تملكهESLAT.

تجدر الملاحظة إلى أن المعادلات للتنبؤ SAT لفظي ورياضيات تظهر أن لديها قابلية تطبيق محدودة بسبب كونها غير قابلة للاستخدام من أجل درجات نهائية للاختبار بدرجات نهائية ESLAT أدنى من 550 في العينة لديها علاقة شاردة مع المتغيرات موضع الاهتمام. كان تفسير هذا بما معناه، أن مستوى معيناً من المقدرة

في اللغة الإنكليزية، كما تم قياسها بواسطة ESLAT يكون مطلوباً قبل أن تصبح الدرجات النهائية متصلة نظامياً بالدرجات النهائية الأخرى للاختبار، وأكثر أهمية، قبل أن تستقر العلاقات بين الدرجات النهائية في تراكيب بنيوية مشابهة جرى قياسها بالإنكليزية و الإسبانية. إن الدور البارز لـ ESLAT في التنبؤ بالدرجات النهائية SAT لفظي حتى في هذه المجموعة من الدرجات النهائية العالية ESLAT (550) هو بالتقريب معيار الانحراف [118] فوق المتوسط 446 في عدد السكان الكامل (PAA) يدفع إلى المقدمة، المشكلات في محاولة ربط الدرجات النهائية في الاختبارات مثل الـ PAA والـ SATI التي يجري إعطاؤها في لغات مختلفة إلى مجموعات من خلفيات ثقافية مختلفة. إن ربط درجات نهائية باستخدام علم منهج التنبؤ قد يبدو أنه يقدم نتائج أكثر قابلية للتفسير منه في المحاولة لتأسيس جدول اتفاق باستخدام علم منهج للتساوي التقليدي للاختبار المعتمد الذي جرى توظيفه في الدراستين الأوليتين.

إن تصميم الدراسة الثالثة لديه بالتأكيد عدد من العوائق التي ينبغي لفت النظر إليها. إن الدراسة هي دراسة تنبؤ مستخدمة لتصميم مجموعة ثنائية اللغة. إن عوائق دراسات التنبؤ وتصاميم مجموعة ثنائيّ اللغة التي جرت الإشارة إليها سابقاً في هذا الفصل، لكن تصميم المجموعة ثنائية اللغة المستخدم في هذه الدراسة هو مختلف عن التصاميم التي جرت مناقشتها سابقاً والتي فيها تم استخدام ESLAT كمتغير ملازم. إن العائق الظاهر لهذا التصميم هو أن معادلات التنبؤ المحصّلة من هذه الدراسة تكون مختصّة بمجموعة والعينة المستخدمة لأجل هذه الدراسة غير ممثلة لجميع المتقدمين للاختبار الآخذين اختبار الـ PAA تألفت من طلاب قدموا في المقام الأول من مدارس عليا خاصة في بورتوريكو بمستويات أعلى في التمكن من اللغة الإنكليزية من تلك المستويات الموجودة إلى حد نموذجي بين طلاب المدارس العليا في بورتوريكو. على أية حال، إنه فقط هذا النمط من المتقدم للاختبار الذي يتطلع على نحو نموذجي إلى تعليم بعد الثانوي في الولايات

المتحدة. وهكذا، مع أن نتائج الدراسة قد لا تصف العلاقة بين الدرجات النهائية المحصّلة في الـ PAA والـ SATI من أجل كل طلاب المدارس العليا في بورتوريكو، ربما تكون النتائج صالحة تماماً من أجل عدد منتقى من الطلاب الذين يعزمون على إكمال دراستهم في الولايات المتحدة.

أشار شميت و آل (1998) إلى أن التعميم إلى مجموعات أخرى أبعد من تلك الممثلة في العينة (تشمل مجموعات تأخذ بعين اختبار الـ PAA في بلدان أمريكا اللاتينية، أخرى غير بورتوريكو)، ربما لا يكون ملائماً نظراً لأن العلاقة بين SATI و PAA و ESLAT ربما تختلف في تلك البلدان الأخرى.

عائق إضافي لعلم المنهج الممثل بهذه الدراسة هو أنه لم ينجم عنه جدول اتفاق يسمح بمقارنات مباشرة لمجموعات فرعية من طلاب يأخذون الـ PAA مع مجموعات فرعية لطلاب يأخذون الـ SATI كنتيجة لهذه الدراسة، تم تزويد مستخدمي الدرجات النهائية بجدول يتطلب إدخالاً مع الدرجة النهائية PAA و ESLAT كلتيهما والدرجة النهائية للقراءة SATI المتنبأ بها من النص المطبوع للجدول. بناء عليه كان الربح في صحة تنبؤات الدرجة النهائية يتوازن مع خسارة للقابلية العملية أو الملائمة لمستخدم الدرجة النهائية.

نظراً لأن تطبيق علم المنهج المستخدم في تطوير العلاقة بين الدرجات النهائية PAA و SATI في الدراسة الثالثة لم تظهر نتائجه في جدول الاتفاق، ليس بالإمكان أن نقارن نتائج هذه الدراسة مع تلك المحصّلة في الدراستين السابقتين المقامة من قبل آنغوف ومودو (1973) وآنغوف وكوك (1988). بالرغم من أنه تجدر الإشارة إلى أنه نظراً لأن كلا الاختبارين PAA و (SAT) قد تم تعديله إلى حد بعيد منذ أن تم إكمال الدراستين السابقتين، قد تكون مقارنة النتائج عبر الدراسات الثلاث موضع تساؤل، حتى إذا أيّد علم المنهج المستخدم لربط الاختبارات في الدراسة الثالثة، تطوير جدول الاتفاق.

مناقشة

بالإمكان تعلم عدد من الدروس المهمة من تقييم العمل الذي تم إجراؤه عبر العشرين سنة الماضية والذي ركّز على ربط الدرجات النهائية المحصّلة في PAA بدرجات نهائية محصّلة في الـ SAT. حاولت كل من الدراسات الثلاث التي تجري مناقشتها هنا تحسين نتائج الدراسات السابقة بتطبيق التفكير الأكثر شيوعاً في نظرية القياس السيكولوجي والتطورات التقنية الأقرب حداثة. مع ذلك عرضت حتى الدراسة الأقرب حداثة، التي أجراها شميت وآل، عدداً من العوائق الجدية. بالتأكيد أن الخبرات المكتسبة من دراسات ربط PAA/SAT الثلاث تعرض بقوة كم يكون صعباً الحصول على درجات نهائية صحيحة وقابلة للمقارنة من اختبارات تم إعطاؤها إلى مجموعات تختلف في اللغة والثقافة.

من المحتمل أن التقدم الأكثر أهمية في دراسة آغنوف - مودو (آنغوف ومودو 1953) كان تطبيق تقنية الرسم البياني المثلى لأجل كشف مفردات في منظومة تعادل المفردة "العامة" التي لم تسلك طريقة متشابهة في المجموعات الناطقة بالإسبانية والناطقة بالإنكليزية. استخدم آغنوف ومودو هذه التقنية الجديدة، التي كانت قد تم تطويرها لغربلة المفردات من أجل انحياز عرقي آنغوف وفورد، (1973) تتضمن سير العملية تعيين قيم صعوبة المفردة (المثلثات) من أجل مفردات جرى تقريرها على المجموعتين موضع الاهتمام، وشطب تلك المفردات التي تقع بعيداً عن المحور الرئيس للقطع الناقص المشكل بالرسم البياني. أدرك آنغوف ومودو بوقت مبكر أن مأزقاً جديداً واحداً في دراسات حضارية متداخلة/لغوية متداخلة كان على الرغم من الترجمة الأكثر شدة في التدقيق والترجمة المعادة، يجب أن يتم إحصائياً غربلة المفردات التي يتوقع أن تتصرف بصورة مشابهة (مثل مفردات "عامة" في تصميم اختبار معتمد، على الأخص تلك المفردات التي لديها عنصر لغوي/ لفظي متين). يظهر بحث أخير لـ ميونز، هامبلتون، وإكسنغ (2001) أيضاً رسوماً بيانية مثلثية لا تزال تستطيع أن تكون مفيدة في كشف المفردات الإشكالية ولو بأحجام عينات صغيرة.

إن الدراسة التي قام بها آنغوف وكوك (1988) مبنية على تصميم دراسة سابقة مع بعض التحسينات التقنية والمنهجية. أمل المؤلفون في أن استخدام إجراءات IRT، لتحل مكان استخدام إجراءات نظرية الاختبار الكلاسيكية، التي كانت تستخدم في الدراسة الأولى، ستقدم نتائج محسنة. في الحقيقة، أثبتت إجراءات IRT لاكتشاف DIF انظر لورد، 1980 كونها إجراءات قوية جداً من أجل غربلة المفردات العامة. كان المؤلفون واثقين جداً من أنه في الوقت الذي يكونون فيه قد أكملوا غربلة المفردات سيكونون قادرين على تشييد اختبار "عام" يمكن استخدامه لربط الأغراض دون خطر تمييز أية مجموعة مركز الاهتمام. على أية حال، بدأ المؤلفون يشكّون في الافتراضات الضمنية للعمل الذي يقومون به. هل كان الاختباران PAA و (SAT) يقيسان تركيبات بنيوية متشابهة بدرجة كافية لتأييد تطوير جدول الاتفاق؟ ماذا كان يعني استخدام درجة نهائية في اختبار PAA لطالب كان يتكلم الإسبانية فقط لتقدير درجة نهائية للطالب في القسم اللفظي من اختبار الـ SAT؟

كنتيجة للأمور المقلقة المثارة من قبل مؤلفي دراسة الربط SAT/PAA الثانية، جرت مراجعة علم المنهج المستخدم في الدراسة الثالثة بشكل كلي. استخدم مؤلفو الدراسة الثالثة شميت وآل، (1998) إجراءات الارتداد لتطوير العلاقة بين الدرجات النهائية PAA و SAT في مجرى تحليل المعطيات للدراسة الثالثة، وجدوا أن الارتباطات بين الدرجات النهائية المحصّلة في الاختبارات PAA اللفظية و SATI اللفظية (لأجل العينة ثنائية اللغة المستخدمة في الدراسة) كانت فقط أعلى قليلاً من الارتباطات بين الدرجات النهائية المحصّلة في الاختبارات PAA اللفظية و PAA الرياضيات بالرغم من أنه، كما جرت الإشارة سابقاً، توجد تقنيات إحصائية يمكن استخدامها لتطوير جداول الاتفاق عندما لا تقيس الاختبارات الشيء نفسه (و في الحقيقة، العمل الذي تم إجراؤه في الدراسة الثانية هو مثال ممتاز لهذا النمط من العمل) يبقى السؤال، كيف يفسّر المرء النتائج من التطبيق لجدول اتفاق مطور تحت تلك الظروف؟

اختار شميت وآل (1998) تطوير معادلات التنبؤ من أجل التنبؤ بالدرجات النهائية SAT من الدرجات النهائية PAA أخذت المعادلات في الحسبان ليس فقط المقدرة اللفظية أو الرياضية للمتقدم للاختبار، كما تم قياسها بـ PAA، لكن أيضاً اعتبرت المقدرة باللغة الإنكليزية للمتقدم للاختبار، كما تم قياسها بـESLAT بالرغم من أن معادلات التنبؤ غير صالحة للاستخدام ولا يمكن استخدامها بجاهزية في المقارنة لمجموعات متقدمين للاختبار، فإنها بالتأكيد تزود بإجابة أكثر دقة على السؤال كيف لطالب يحصل على درجات نهائية بمستوى معين في الـPAA أن يحصل على درجات نهائية في الـ SATI.

إن الأسئلة التي تبقى ليجري اكتشافها عند اعتبار ربط PAA و SATI هي: الوصف للتشابه أو الفروق بين التركيبات البنيوية المقاسة من الـ PAA و SATI وكيف تأثرت هذه التشابهات أو الفروق بالمقدرة في اللغة الإنكليزية. أضف إلى ذلك أنه من المهم أن نبقي في الذهن أن الكليات لا تهتم كثيراً بتنبؤات الدرجات النهائية SATI من درجة نهائية PAA بقدر ما تهتم في اتخاذ قرارات صالحة حول كم سيكون الطلاب ناجحين إذا تم قبولهم في كلية معينة. إن العلاقة بين الدرجات النهائية PAA وSATI، والتي جرى تطويرها في الدراسة SATI/PAA الثالثة، لا تتطلب مصادقة رسمية باختبار العلاقة بين الدرجات النهائية SATI المتنبأ بها والأداء في الكلية، كما جرى في "متوسط درجة طالب الصف الأول الجامعي"، أو في معيار آخر ما ذو أهمية.

من المهم أن نبقي النتائج والدروس المتعلّمة من الدراسات SAT/PAA الثلاث في أذهاننا عند مراجعة عمل مشاريع أخرى لتكييف الاختبار. ليس تكييف الاختبارات عملية تافهة، بل إنها عملية مهمة جداً تؤثر بدرجة عظيمة على صحة الدرجات النهائية للاختبار. استناداً إلى الاستخدام اللا متناه للدرجات النهائية للاختبار، يمكن أن يكون من الأهمية بمكان للطلاب، والموظفين، وسكان بلدان أخرى

أن يجري إعطاؤهم الفرصة لعرض مهاراتهم وقدراتهم في اختبارات يتم إعطاؤها بلغتهم الأصلية. درسان هامان تم تعلمهما من الدراسات SAT/PAA الثلاث يشيران إلى أنه:

1- من المهم أن نأخذ بالحسبان كيف سيجري استخدام وتفسير الدرجات النهائية للاختبار. سيعتمد بدرجة كبيرة تصميم دراسة الربط والنماذج المختارة على الاستخدامات الممكنة وتفسيرات الدرجات النهائية للاختبار.

2- لا توجد طريقة بسيطة للقيام بوظيفة ذات نوعية عالية لتكييف الاختبارات من أجل لغات وثقافات مختلفة. الاختبارات المكيّفة هي عملية مجهدة تتطلب تنفيذاً حذراً، ليس فقط انتباهاً دقيقاً فقط للعملية التطويرية للاختبار، بل يكون الانتباه لعملية إدارة الاختبار وتفسير الدرجات النهائية على قدر مساو من الأهمية.

لحسن الحظ يجري القيام بعمل جيد في موضوع تكييف الاختبار وبالاهتمام المتزايد في المجال. ستكون الحلول لما قد يبدو مشكلات متفاعلة على الأغلب متوفرة لدينا في المستقبل.

شكر

يقدر المؤلفون الإسهامات لـ أنتوني ماغرينا، فيل دورانز، ودانيل إيغنور في الإعداد لهذا الفصل.

*****

المراجع

Alderman, D. L. (1981). *Language proficiency as a moderator variable in testing academic aptitude* (TOEFL Research Rep. No. 10, RR81-41). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Rep. No. 88-2). New York: College Entrance Examination Board.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–106.

Angoff, W. H., & Modu, C. C. (1973). *Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Rep. No. 3). New York: College Entrance Examination Board.

Boldt, R. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade County high school volunteers* (Statistical Rep. No. 69-31). Princeton, NJ: Educational Testing Service.

The College Board. (1995). *Cambiod en el examen de admision del College Board: La nueva PAA* [Changes in the College Board admissions tests: The new PAA]. San Juan: Oficina de Puerto Rico y de Actividades Latinamericanas.

Cook, L. L. (1995, April). *Lessons learned: Implementing change in the SAT*. Paper presented at the meeting of the National Council on Educational Measurement, San Francisco.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1989, April). *Equating achievement tests using samples matched on ability*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3*, 37–55.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Transition and adaption issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.

Hambleton, R. K. (1993). Adapting achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*, 57–68.

Hambleton, R. K. (1996, April). *Guidelines for adapting educational and psychological tests*. Paper presented at the meeting of the National Council on Educational Measurement, New York.

Kolen, M. J. (1990). Does matching in an equating work? A discussion. *Applied Measurement in Education, 3*, 97–104.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. 23). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73–95.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in test translation. *International Journal of Testing, 1,* 115–135.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Pennock-Román, M. (1995). *Measuring developed academic abilities using Spanish vs English-langauge tests: PAEG/GRE relationships for Puerto Ricans who are more proficient in Spanish than in English* (GRE Research Rep. No. 89-01). Princeton, NJ: Educational Testing Service.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737–756.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Schmitt, A. P., Dorans, N. J., Magrina, A. & Cook, L. L. (1998). *Predicting scores on the English Language SAT from the Spanish Language PAA and the Spanish Language English as a Second Language Achievement Test.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Sireci, S. G. (1997). Technical issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16,* 12–19.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 147–165.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3,* 105–113.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Terman, L. M. (1916). *The measurement of intelligence.* Boston: Houghton Mifflin.

van de Vijver, F. J. R. & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht, Netherlands: Kluwer Academic.

7

تكييف اختبارات الإنجاز والأهلية
مراجعة موضوعات منهجية

ليندا ل. كوك
اليسيا. شميدت - كاسكالار
الجمعية الدولية للتقويم، بروكسل، بلجيكا

كاثرين براون
خدمات الاختبارات التربوية

كان الاهتمام بتكييف الاختبارات التي تم إنجازها لاستخدامها للغة وثقافة إلى استخدامها مع مجموعة ذات لغة وثقافة مختلفة سائداً بين الباحثين التربويين والنفسيين في القرن العشرين. على سبيل المثال، درس هاميلتون وبولورك (1991) الترجمات الأولى لاختبار "بينت - سيمون" للذكاء. أشاروا إلى أن الاختبار قد ترجم من الفرنسية إلى الإنجليزية في 1911 واستخدم لتقويم ذكاء الدارسين في دار المعلمين في نيوجيرسي (الولايات المتحدة). أشاروا أنه حتى عام 1916 كان قد تم ترجمة الاختبار إلى سبع لغات مختلفة (ستانلي وهوبكنز، 1972). كما أشاروا إلى أن اختبارات ذكاء أخرى ومقاييس مرافقة تم ترجمتها إلى لغة الطلاب الأولى.

أشارت مطبوعات فان دي فيفر (2002) فان دي فيفر ولونر (1995)، فان دي فيفر ولونغ (1997، 2000) إلى أنه تضاعفت حديثاً مطبوعات تعالج الاختلافات

والمقارنات عبر الثقافات، وعزوا ازدياد تلك المطبوعات إلى عدة أسباب مثل العالمية في الاقتصاد، السياحة، الهجرة المستمرة وتغيرات سياسية.

إن الغرض من هذا الفصل هو مراجعة الموضوعات المنهجية التي ترافق تكييف مقاييس الإنجاز والأهلية. يتكون هذا الفصل من ستة أقسام. يقدم القسم الأول نظرة عامة عن موضوعات التحيز والتكافؤ وعلاقتهما مع مقارنة تقويم النتائج عبر اللغات/ الثقافات. يتبع تلك النظرة العامة أقسام متفرقة تركز على موضوعات متعلقة مثل: (أ) تكافؤ البنية، (ب) تكافؤ وحدات القياس، (ج) ترجمة الاختبارات ومواد الاختبارات، (د) تفسير الدرجات واستخدام الاختبار (د) إدارة الاختبار.

نظرة عامة على الموضوعات الرئيسة

أشار فان دي فيفر ولونغ (1997) أن "الأفكار المتكررة عن التقويم في اللغات المتعددة هو التساؤل عن مدى إمكانية استخدام الوسائل التربوية المطورة في الدول الغربية من محيط ثقافات مختلفة (صفحة 61). وضعوا لائحة تتضمن أربعة من تساؤلات المهتمين بتقويم الثقافات المتعددة:

- هل يوفر الاختبار تغطية كاملة للبنية النفسية نفسها للمجموعات الثقافية التي تستخدمه؟
- هل من الممكن استخدام قواعد إدارية واحدة أو يجب تكييفها؟
- كيف تستطيع التعامل مع المتغيرات الكثيرة في مهارات اللغة الأم واللغة المضيفة لمجموعات مهاجرة؟
- هل يوجد بدائل أخرى عندما تبين أن الاختبارات الغربية غير مناسبة؟

إن تلك الأسئلة الأربعة لها علاقة بمواقف تحصل عند تكييف اختبارات الأهلية والإنجاز لاستخدامها عبر الثقافات/ اللغات المختلفة.

إن معظم المطبوعات عن التقويم عبر الثقافات خصصت لاختبار وتقويم مقارنة بين الاختبار عبر ثقافات ولغات مختلفة.

إن التساؤلات عن تأثير الإجراءات الإدارية بالإضافة إلى الاختلافات اللغوية مهمة بشكل خاص لتكييف اختبار الأهلية والإنجاز ملائم ولإثبات صدق مقارنة درجات الاختبار عبر الثقافات.

وضح فان دي فيفر وتانزر (1997) ثلاثة من أنواع التحيز التي يمكن أن تؤثر في التقويم عبر الثقافات، وعبر اللغات، التحيز في البنية، التحيز في المنهج والتحيز في بنود الاختبار. وقد ناقشوا أيضاً تأثير نوعين من التكافؤ على الأبحاث والتقويم عبر الثقافات: تكافؤ البنية وتكافؤ وحدة القياس.

إن النقطة التي أثارها فان دي فيفر وتانزر هي أن السؤال الأساسي في كل الدراسات عبر الثقافات هو عما إذا كانت الدرجات الموجودة التي أعطيت لمجموعات مختلفة يمكن تفسيرها بطريقة واحدة. وقد أكدوا أن موضوع التحيز والتكافؤ أساسيان في تلك النقطة.

عرَّف فان دي فيفر وتانزر (1997) تحيز البنية بأنها تحدث إذا كانت البنية المقاسة غير متماثلة عبر المجموعات الثقافية، وقد استعملوا كمثال اختبارات الذكاء الغربية. اعتبر هؤلاء الكتاب بشكل أساسي أن تحيز البنية ليس مصطلحاً ينطبق على أداة ما ولكنه ينطبق على خصائص مقارنة عبر الثقافات، أشاروا إلى أن الأداة التي تكشف التحيز في مقارنة موضوعات بين اليابانية والألمانية قد لا تكشف التحيز في مقارنة بين الألمانية والهولندية.

حسب فان دي فيفر وتاتزر (1997) فإن التحيز في العينات ينتج عن اختلافات في العينات المتعلقة بالمتغيرات (غير متغيرات الاهتمام) التي يمكن أن تؤثر في نتائج البحث. وقد أعطوا مثال المجموعات الثقافية التي تختلف في الخلفية الثقافية والحوافز. إذا لم تكن أي منهما واحدة في متغيرات الاهتمام فإنهم بالتأكيد

سيحدثون ارتفاعاً في مقارنة تلك المتغيرات التي هي مركز الاهتمام في تلك الدراسة.

استخدم فان دي فيفر وتانزر (1997) في مناقشة تحيز الأداة "الإثارة المألوفة" كمثال معروف جداً. وقد توسعوا في ذلك المثال بالاستشهاد بأعمال هوي وترياندیس (1989) اللذين وجدا أن الأمريكيين من أصول لاتينية يميلون إلى اختيار الحد الأقصى في مدرج قياس الشخصية ذي خمس درجات بنسبة أكبر من الأمريكيين البيض. اكتشف هوي وترياندیس أن الارتباك الحاصل بسبب تفضيل الخيار قد اختفى عندما استعمل مقياس ذو عشر درجات.

إن النوع الثالث لتحيز المنهج الذي بحثه فان دي فيفر وتانزر (1997) هو التحيز الإداري. اعتبر بعض الكتاب أيضاً موضوع التحيز الإداري كتهديد لصدق تفسير درجات الاختبارات المكيفة. على سبيل المثال ناقش كيسنجر (1994) حقيقة أن الاختلافات الثقافية لمجموعات قومية تتغير حسب مستوى معرفتهم ببنود بنية الاختبار المختلفة واقترح استخدام عدد كاف من التمارين العملية لجعل البنية الجديد مألوفاً إلى الطلاب الممتحنين.

أخيراً، بالإضافة إلى مناقشة تحيز الأسلوب والبنية فقد ناقش فان دي فيفر وتانزر (1997) تحيز البند كنوع مهم الذي يسبب الإرباك في دراسات عبر الثقافات. وقد وضعوا لائحة بالأسباب التالية لتحيز البند في تلك الدراسات: "ترجمة سيئة للبند" بنود غامضة، عوامل مزعجة (قد يسبب البند آثاراً إضافية أو قدرات)، مواصفات ثقافية (اختلافات ثانوية في معنى المفاهيم/ ملاءمة محتوى البند) (صفحة 268).

من غير الضروري القول إن موضوعات تحيز البند تُعد تحدياً هائلاً إلى صدق أي مقارنة أهلية عبر الثقافات أو درجات اختبار إنجازات. إن الموضوعات المتعلقة بتحيز البند، مثل ترجمة الاختبار والإجراءات المتبعة في كشف التفاوت الوظيفي

للبند (DIF) قد نوقشت في الفصل الرابع في هذا الكتاب؛ ولذلك لن تؤخذ بعين الاعتبار أبعد من ذلك.

في مناقشة التكافؤ في دراسات عبر الثقافات ميز فان دي فيفر وتانزر (1997) بين التكافؤ في البنية والتكافؤ بين وحدات القياس. كان تعريفهم لتكافؤ البنية أن البنية تقاس عبر كل الثقافات للمجموعات المعينة بصرف النظر عما إذا كانت أدوات القياس المستخدمة متماثلة. أضافوا أنه من الممكن للأداة نفسها قياس بنيات مختلفة في ثقافات مختلفة، أو أن البنية المقاسة بالأداة نفسها قد تتشابه جزئياً بين الثقافتين.

عرف فان دي فيفر وتانزر (1997) نوعاً ثانياً من التكافؤ له ذات أهمية تكافؤ البند، تكافؤ وحدة القياس يحصل ذلك الموقف إذا كان القياسان لهم ذات وحدة القياس مع اختلاف المصدر. وقد وضحوا ذلك التعريف بالقول: "بكلمات أخرى إن مقياس القياس الأول قد تم تحويله بتساوٍ مستمر بالمقارنة مع المقياس الآخر" (صفحة 266). وقد أشاروا إلى أن درجات المقاييس التي لها تلك المواصفات لا يمكن مقارنتها مباشرة، ولكن إذا عرف عامل التساوي فإنه من الممكن تعديل الدرجات لجعلها مناسبة للمقارنة.

أخيراً، إذا أراد أحد ما أن يقوم بمقارنة جيدة لدرجات اختبارات الإنجاز أو المهارات، يجب عليه القيام بما أشار إليه فان دي فيفر وتانزر (1997) وهو مقارنة مدرجة، وقد جرى تعريفها بأنها المستوى الأعلى من مستوى تكافؤ وحدة القياس لاثنين من المقاييس. أشاروا أن ذلك النوع من التكافؤ يمكن الحصول عليه عندما يكون المقياسان لهما ذات المصدر ووحدة القياس. أضافوا بالقول إن التكافؤ المدرج هو مطلب أساسي للقيام بمقارنة تقويم النتائج عبر التقانات. وقد أكدوا على أن أي شكل من التحيز، الأسلوب، البنية إلى ما هنالك سوف يؤدي إلى اختلال وإضعاف تكافؤ وحدات القياس.

تكافؤ البنية

يقدم هذا القسم من الفصل مناقشة أهمية تكافؤ البنية في الدراسات عبر الثقافات كما يقدم نظرة عامة لإجراءات منتقاة تستخدم لتقويم تكافؤ البنية عبر مجموعات مختلفة.

أهمية تكافؤ البنية

إن كثيراً من الدراسات النظرية والعملية سائدة في الأعمال المنشورة حول التقويم عبر الثقافات. كثير من الكتاب مثل بورتينغا (1989، 1983)، فان دي فيفر ولونغ (1997)، فان دي فيفر وتانزر (1997) قد طورا تعريفاً بليغاً ومناقشات لتكافؤ البنية.

على ذلك الخط ناقش بورتينغا (1989) ما أشار إليه "منطق المقارنة". وقد أصر على أن المقارنة بين الأفراد، أو المجموعات يمكن أن تكون مضللة لسببين. السبب الأول هو أن الموقف الذي استخدم في المقارنة قد لا يكون متشابهاً عبر الأفراد أو المجموعات (تكافؤ البنية). وقد أعطى مثال، المقارنة بين الطول والوزن، أي أنه لا معنى له أن يكون شخص أطول من شخص آخر كبير الوزن. أما السبب الثاني فهو أن وحدات القياس قد لا تكون متشابهة (تكافؤ وحدة القياس)، على سبيل المثال أن الطول المقاس بالبوصه لا يمكن مقارنته بالطول المقاس بالسنتيمتر.

توسع بورتينغا حول العلاقة بين تكافؤ البنية وتكافؤ وحدة القياس بالإشارة إلى أن إمكانية تشكيل مقياس قياس يمكن مقارنته لنسختي اختبار مختلفتين في المحتوى ضئيلة جداً، بكلمات أخرى من الواضح أن تكافؤ البنية شرط أساسي لحدوث تكافؤ القياس.

تحرى بورتينغا (1983) في ورقة بحث ثانية بتوسع العلاقة بين تكافؤ البنية وأشكال أخرى من التكافؤ وقد ناقش مضمون تحليل يظهر الاختلافات في تقويم النتائج عبر المجموعات. أشار إلى أنه ترك للباحث تقرير ما إذا كان فقدان نتائج

قابلة للمقارنة ينتج عن بنية غير متكافئة، وحدات قياس غير متكافئة أو عن اختلافات حقيقية ضمن المجموعات.

استمر بورتينغا (1983) بالقول إن تحليل مقارنة نتائج الاختبار تؤدي غالباً إلى أحد حقول صحة البنية. وقد اعترف بنظرية أن تكافؤ البنية وصدق البنية كثيراً ما تتشابك إلى حد كبير. كانت النقطة التي أثارها رئيسة في تحليل تكافؤ البنية، كان السؤال الرئيس هو عما إذا جرى قياس البنية ذاتها، بينما في صدق البنية كان الموضوع الرئيس هو أي بنية قد جرى قياسها.

أكد بورتينغا على أهمية البحث المستند على النظريات. شدد على أن "تحليل مقارنة ذات معنى يحتاج إلى إطار عمل نظري كأساس يستطيع التعهد بصراحة أية علاقة بين أية متغيرات يمكن أن تكون ثابتة على المجموعات" (صفحة 246). أشار بورتينغا كمثال إلى كتابات فان دي فيفر ودرنث (1980). حسب فان دي فيفر ودرنث فإن مناقشة قياس الطول والوزن تقدم نتائج مقارنة عبر الثقافات حتى إذا كانت العلاقة بين تلك المتغيرات قد تختلف عبر المجموعات. على كل حال أشاروا إلى أنه إذا كان الطول والوزن سيستخدم لتخمين المتغير الثالث، على سبيل المثال محيط الخصر، عندئذ لكي يجري اعتبار لقياس متكافئ عبر المجموعات التي يجري دراستها، يجب أن يكون للطول، للوزن، لمحيط الخصر علاقة متماثلة عبر تلك المجموعات. كنتيجة لذلك فإن من الصعب تقويم التكافؤ أو صدق القياس عبر المجموعات التي يجري دراستها دون نظرية مدروسة عن العلاقة بين الطول، الوزن ومحيط الخصر.

عرف فان دي فيفر وتانزر (1997) تكافؤ البنية بأنه "البنية التي يجري قياسه عبر كل المجموعات الثقافية التي تُدرس بغض النظر عما إذا كان قياس البند يستند على أدوات متماثلة عبر الثقافات أم لا" (صفحة 265). أشاروا إلى أن تكافؤ البنية يمكن أن يحدث عندما تتشابك البنيات عبر الثقافات جزئياً فقط أو عندما تترافق

البنيات مع سلوكيات أو خصائص مختلفة كنتيجة لاختلاف الثقافات وقد قدموا ملخصاً عن مصادر تحيز البنية التالي (عدم وجود التكافؤ): "تشابك جزئي فقط في البنية عبر الثقافات؛ تفاوت في ملائمة السلوك المتعلق بالبنية (مهارات لا تنتمي إلى الذخيرة الخلفية لإحدى المجموعات الثقافية)؛ نماذج سيئة لكل السلوكيات المرافقة (وسائل إيضاح قصيرة)؛ تغطية غير كاملة لكل النواحي المتعلقة بالبنية (عدم وجود نماذج لكل الميادين التعليمية) (صفحة 268).

استطرد الكتاب بتفصيل الاستراتيجية القيمة التالية لتعريف موضوعات عن تقويم تحيز البنية (أو التكافؤ) عبر الثقافات:

- اللامركزية (مثال: تطوير الأدوات الايضاحية في عدة ثقافات في وقت واحد).
- طريقة المقاربة (مثال: استقلال تطور الأداة ضمن الثقافة وتطبيق كل الأدوات عبر الثقافات لاحقاً).
- استخدام خبراء لتقديم المعلومات اللغوية عن اللغة والثقافة المحلية.
- استخدام مسح محلي (مثال: تحليل محتوى الأسئلة ذات إجابة حرة).
- تطبيق أداة غير مألوفة (التفكير بصوت عال).
- مقارنة شبكات مرتبطة مع بعضها في علم المنطق (مثال، تقارب/ صحة تميز الأبحاث، أبحاث في طرق عديدة لها سمة وحيدة، مفاهيم عبارات رئيسة) (صفحة، 272).

فيما يلي مناقشة المزيد في الاستراتيجيات الشائعة لتقويم تكافؤ البنية التي جرى تعريفها سابقاً، مع نماذج منتقاة لتطبيقات ناجمة.

تقييم تكافؤ البنية

وصف هيوي وترياندس (1985) طرقاً عديدة استخدمت لبرهان تكافؤ البنية واعتبروا استخدام طرق متعددة أساسياً في تلك المسألة. بدؤوا بوصف قواعد

تراجعية كطريق لبرهان التكافؤ عبر الثقافات. أكدوا أنه من الدقة الاستنتاج أن التقييم المتكافئ عبر المجموعات يتعلق بطريقة مماثلة بمعايير خارجية. استطردوا بالقول إن هذه طريقة بسيطة واقتصادية لمسألة التكافؤ، لكنهم أشاروا إلى أن الاختلافات في متغيرات النماذج وجدارة الأدوات يمكن أن تسبب عدم استقرار تردي تقدير الدرجات الذي هو "في الأساس إنذار كاذب".

ناقش هيوي وتريانديس (1985) أيضاً نظرية الإجابة للبند (IRT)، طرق لإقامة التكافؤ (انظير سيرغي والالوف، 2003). وقد بينوا أن الأداة ذات منحى مواصفات للبند (ICCs) عبر الثقافات قد برهنت، ولو بشكل جزئي، على تكافؤ بنودها (وكنتيجة لتكافؤ البنية) وتكافؤ المدرج" (صفحة 139).

ناقش آنوف وكوك استخدام نظرية الإجابة للبند (IRT) لتطوير سلسلة من بنود تكافؤ البنية التي تستخدم لربط اختبارات قدرات متطورة أعطيت لمتحنين ناطقين في اللغة الإنجليزية وآخرين ناطقين باللغة الإسبانية. استنتجوا أن طريقة (IRT) كانت فعالة عند الاستخدام لذلك الهدف. كتب بالتفصيل عن بحث آنجوف وكوك في الفصل السادس.

بالإضافة إلى نقاش حول طريقة تردي التقدير وطريقة (IRT)، ناقش هيوي وتياندیس أيضاً تطابق البنية. أشاروا إلى حقيقة أنه إذا أُريد اعتبار البنية متكافئة عبر الثقافات يجب أن تكون البنيات الداخلية والعلاقة بينهم متشابهة عبر الثقافات والاهتمامات. اقترحوا عوامل تحليلية ومعايير متعددة الأبعاد لتقنية إحصائية يمكن أن تكون مفيدة للاستخدام لفهم كيفية عمل البنية عبر ثقافات مختلفة.

استخدم سيرغي، فتزجيرالد، وزينغ (1998) مجموعة مشتركة من تحليل قواعد العناصر، تحليل العناصر التأكيدية، معايير متعددة الأبعاد لتقييم تكافؤ البنية لاختبارات أجريت عبر الانترنت للمتحنين إنجليز، فرنسيين، ألمانيين ويابانيين في لغاتهم الأم أجرى الباحثون تحليل قواعد العناصر على مستوى المعطيات،

مستوى البند الواحد ومستوى مجموعة البنود. (ينصح استخدام مجموعة من البنود في عملية تحليل العناصر لتفادي صعوبات غير منطقية التي من الممكن حدوثها إذا كان الارتباط على مستوى معطيات البند الواحد). يجري تحليل المعايير المتعددة الأبعاد على معطيات مجموعة بنود فقط، أشار عدد من الباحثين إلى إمكانية استخدام إجراءات تحليل العناصر التأكيدية لتقديم تكافؤ البند غير المجموعات (انظر. غيرل، 2000) تقصى أفرسون، غيريرو وليتوسيس تكافؤ بنية SATI، اختبار الرياضيات (SAT-M) وقسم الرياضيات في اختبار الأهلية (PAA-M) الذي يجري لطلاب مدارس ذوي لغتين في بورتوريكو. استخدم الباحثون كلاً من تقنية تحليل العناصر التأكيدية وتحليل العناصر التمهيدية. استخدم أفرسون وزملاؤه مجموعة تقنيات لتطوير المعلومات على تحليل العناصر التمهيدية.

تكافؤ وحدات القياس

إن إقامة وحدة قياس مشتركة للدرجات التي تم الحصول عليها من اختبارات الإنجاز واختبارات الأهلية التي أعطيت بلغات مختلفة للممتحنين ذوي خلفيات ثقافية مختلفة عسيرة جداً لأسباب عديدة جرى ذكرها في أوائل هذا الفصل، إن السبب الذي يجعل ذلك مشكلة جديدة هو أن أحد المفاهيم المفترضة لأكثر الطرق المستخدمة في إقامة وحدة قياس مشتركة (طرق المقاييس المترابطة) هو أن التقويمات التي ستربط ببعضها تقيس ذات البنيان أو بنيات متشابهة، يحتوي الفصل السادس نظرة عامة ومناقشة إجراءات وموضوعات متعلقة بإقامة تكافؤ بنود القياس.

ترجمة الاختبارات ومواد الاختبارات

تستخدم دراسات مقارنة الإنجاز أو الأهلية في لغات مختلفة عبر الثقافات أدوات توضيحية جرى ترجمتها وتكييفها إلى اللغة والثقافة المستهدفة. لا تتضمن عملية تطوير أدوات متكافئة في أكثر من لغة ترجمة بنود الاختبار ومواد الاختبار

ولكنه يتضمن تغيرات في هيكلية البند وإجراءات الاختبار أيضاً (تكييف الاختبار). يجب أن نأخذ بعين الاعتبار الموضوعات المتعددة التي لها علاقة بترجمة الاختبار للحصول على أدوات مناسبة للمقارنة عبر الثقافات. "يجب أن لا تعكس الترجمة الجيدة معنى البند الأصلي، لكنها يجب أن تحاول الحفاظ على صلة الاهتمام العقلي والألفة مع محتوى البند وإلا قد يتغير ما يرغب البند في قياسه" (اريكان، 1999، صفحة 2). شكلت الهيئة الدولية للاختبارات عام 1992 مجلساً مكوناً من 13 عضواً لتطوير معيار فني/ تقني لتطوير الاختبار: قدمت الهيئة مجموعة من 22 خطاً من خطوط رئيسة لتكييف اختبارات ثقافية ونفسية (هامبلتون، 1994، انظر الفصل الأول من هذا الكتاب). تناقش سبعة من تلك المجموعة بشكل خاص موضوعات متعلقة بعملية الترجمة وتأثيرها على صلاحة الأدوات التي طورت لإقامة مقارنة الإنجاز والأهلية عبر اللغات/ الثقافات بغية تنظيم ملخص عن الموضوعات المتعلقة بترجمة الاختبار، يجري تقديم كل من تلك الخطوط العريضة مع تفسير ملخص ومراجع وثيقة الصلة بتلك الموضوعات.

1- يجب على مطوري /ناشري Developers/ Publishers الأدوات التأكد أن عملية التكيف تأخذ بعين الاعتبار الاختلافات اللغوية/ الثقافية للمجموعات المقصودة لاستخدم تلك النسخ المكيفة للأداة. إن وجود مترجمين ذوي خبرة وكفاءة مهم جداً لعملية ترجمة الاختبار؛ لأن عملهم له تأثير رئيس على صدق وجدارة درجات الاختبار. تؤكد تلك الخطة على ذلك لأن الخبرة باللغتين (أو أكثر) المعرفة والخبرة في الثقافات، مضمون الاختبار وقواعد القياس أساسية ويجب استخدام فريق من الخبراء، كان هناك خطأ عام وهو استخدام أشخاص متوفرين غير مؤهلين. وحتى عند استخدام شركات للترجمة وتحديد شروط العقد على أن خبراء ثنائيي اللغة ذوي معرفة بموضوعات الاختبار يجب أن يقوموا بعملية ترجمة البنود، لا تزال هناك فروقات في الترجمة. وصل "ريكارس وكنس" (1999)

إلى قرار وهو أن الحل لتلك الفروقات والحصول على بنود تقنية أكثر لاختبار مؤهلات يجب استخدام خبراء ثنائيي اللغة لمراجعة معاني بنود مضمون الاختبار" (صفحة 16) في دراستهم عن دقة ترجمة "الاختبارات التقنية المترجمة والمقدمة طلب "ريكاس وكنس" أن يكون المترجمون ناطقين باللغة المستهدفة. وقد أوصى باستخدام المترجمين الذين لهم سيطرة على اللغة المستهدفة "ودكوك" (1985) و"هامبلتون" (1993) وبذلك تصبح الترجمة طبيعية وفعالة. إن دراسة ثنائية اللغة قد أظهرت أنه من الأسهل لأشخاص يسيطرون على اللغة المستهدفة التعرف على كلمة في لغة المصدر وتذكر بشكل فعال الكلمة المرادفة في اللغة المستهدفة وليس العكس صحيحاً (بيريز، 1975).

2- يجب على مطوري/ ناشري الأدوات توفير الدليل على أن اللغة المستعملة في الإرشادات، التعليمات والبنود ذاتها وفي كتيب الاختيار مناسبة لكل المجموعات اللغوية والثقافية المقصودة بهذه الأدوات. يمكن أن يكون الاختبار في لغة المصدر معقداً دون أي ضرورة وهذا قد يجعل الترجمة الدقيقة صعبة أو أن يكون لها مفهوم، تعابير وأفكار ليس لها مماثل في اللغة المستهدفة (هامبلتون، 1993، اريكان، 1998، 2000). يجب أن يكون مستوى صعوبة المفردات، القراءة، استخدام القواعد، أسلوب الكتابة والتنقيط متشابه عبر اللغات. إحدى الطرق لتقليص الفروقات في صعوبة المفردات هو استخدام لوائح دائمة / متكررة من المفردات (هامبلتون وكانجي، 1993).

إن المشكلة ليست فقط في عدم وجود لوائح مفردات لغوية ولكن في حال وجودها لا تكون في اللغة المستهدفة بالتحديد. أما في حال الترجمة التقنية ينصح "باريكاس وكونس" (1999) باستخدام ملحق للمصطلحات

التقنية عندما لا يكون هنالك مفردات أو تعابير خاصة في اللغة المستهدفة فإنه من الضروري إضافة بنود ثقافية ذات أهمية محددة إلى أداة المصدر (اللاتركيزية). إن المرونة في تلك الإضافة أو تغير أداة المصدر لإقامة التكافؤ عبر اللغات ممكن عندما تكون أدوات اللغتين المصدر والمستهدفة قد جرى تطويرها في وقت متزامن (هامبلتون وكانجي، 1995).

3- يجب على المطورين/ الناشرين توفير الدليل على أن محتويات البند والمواد المحفزة مألوفة لكل المجموعات المقصودة. عندما يجري تطوير الأدوات مع توقع ترجمتها إلى لغات أخرى واستخدامه في ثقافة ثانية، فمن الضروري تجنب استخدام وحدات قياس، عملات نقود أو مواد محظورة (جداول، أرقام، أشكال) التي قد تؤثر بشكل متفاوت على أداء مجموعات مختلفة. يجب أن تؤخذ تلك المسببات الممكنة التي تساعد على أداء مناسب بعين الاعتبار في مرحلة تطوير الأداة. أوصى هامبلتون وكانجي (1995) بتجنب استخدام وحدات القياس، مثل البوصة أو القدم إلى ما هنالك، لأنها تختلف من بلد إلى آخر.

4- يجب على المطورين/ الناشرين تحقيق دليل عقلاني، لغوي ونفسي، لتحسين دقة عملية التكييف وجمع الأدلة عن تكافؤ ذلك في جميع النسخ المختلفة اللغات كما يجب استخدام الطرق العقلانية قبل استخدام الأداة وتقويمها بشكل إحصائي. إن أكثر الطرق العقلانية شيوعاً لإقامة تكافؤ الترجمة هي الترجمة المبكرة والترجمة الراجعة. إن خطة الترجمة المبكرة هي أن يقوم فريق واحد من المترجمين بترجمة الأداة من لغة المصدر إلى اللغة المستهدفة بينما يقوم فريق بآخر بمراجعة تكافؤ الترجمة مع لغة المصدر. أما الترجمة الراجعة فهي إعادة ترجمة الأداة إلى لغة المصدر من قبل مجموعات مختلفة من المترجمين وتقويم تكافؤها مع الأداة الأصلية من مثل

خبير أو أكثر. مع أن الترجمة الراجعة قد جرى استخدامها بشكل مكثف (برسلين 1970) فإن هامبلتون وباتسولا (1998) يؤكدون أن تلك الطريقة لا تقيّم تشابه بنية الاختبار المقاسة في اللغتين بشكل مباشر، "إن العالم غير المساعد الرئيس في تلك الخطة هو، تقييم تكافؤ الاختيار يجرى في لغة المصدر فقط" (هامبلتون وكانجي، 1995، صفحة 151).

بالرغم من أن هذه الخطة توفر مراجعة أولية لتكافؤ الترجمة، "فإن هناك دليلاً قليلاً لإقامة الدليل على أن المترجمين والخبراء لهم المقدرة على تنبؤ تكافؤ النسخ المتعددة للأداة من مراجعة عامة حتى لو جرت بعناية (هامبلتون وكانجي، 1993، صفحة 15).

5- يجب على المطورين / الناشرين التأكد أن جمع المعطيات يسمح باستخدام تقنيات إحصائية مناسبة لإقامة تكافؤ البند في أداة النسخ المتعددة اللغات، يجب توفر مقدار كاف من النماذج في كل من اللغتين، المصدر والمستهدفة، لإقامة تقنية إحصائية (مقياس متعدد الأبعاد، تحليل العوامل) مفيدة وقد وصف هامبلتون (1993) ثلاثة خطط لجمع المعطيات:

- يجري طلاب ثنائيو اللغة اختباراً باللغتين (المصدر والمستهدفة).
- يجري طلاب أحاديو اللغة الاختبار الأصلي واختبار الترجمة الراجعة.
- يجري طلاب لغة المصدر اختبار في لغة المصدر ويجري طلاب اللغة المستهدفة اختبار في اللغة المستهدفة.

أشار هامبلتون (1993) إلى أن الخطط التي تحتاج إلى أشخاص ثنائيي اللغة صعبة التحقيق؛ لأنه ليس سهلاً إيجاد أشخاص ثنائيي اللغة بارعين في كلتا اللغتين. وأضاف أن الدليل الذي تم جمعه باستخدام نماذج من ثنائي اللغة لا يمكن تعميمه على مجموعة أحاديي اللغة المقصودة. وقد

انتقد خططاً تتضمن أشخاصاً أحاديي اللغة يقومون بنسخ أداة اختبار في لغة المصدر والترجمة الراجعة. أشار إلى أن أحد عيوب ذلك النوع من المقارنة هو أن أداة المصدر وأداة الترجمة الراجعة قد تظهر متشابهة بالرغم من الترجمة السيئة. من الممكن حدوث ذلك إذا استخدم المترجمون مجموعة مشتركة من قواعد الترجمة أو إذا احتفظت الترجمة بنواح غير مناسبة من لغة المصدر مثل بنية القواعد الواحدة إلى ما هنالك. فصّل هامبلتون الخطة التي يجري فيها طلاب مجموعة أحاديي اللغة اختبار أداة لغة المصدر وطلاب مجموعة أخرى أحاديي اللغة اختبار اللغة المستهدفة ويتم "ترابط" الاختبارين بواساطة مجموعة من البنود. إن ميزة تلك الخطة هي أن نماذج المجموعتين، المصدر والمستهدفة، تستخدم في التحليل وأن النتائج تعمم غالباً في كل المجموعات المشاركة.

6- يجب على المطورين/ الناشرين استعمال تقنيات إحصائية مناسبة كي (أ) إقامة تكافؤ النسخة المختلفة للأداة و(ب) التعرف على صعوبات في عناصر الأداة والتي يمكن أن تكون غير ملائمة لواحدة أو أكثر في المجموعات المقصودة. كإضافة إلى التقنية العقلانية يمكن استخدام الطرق الإحصائية للتأكد من ملاءمة ترجمة الاختبار. نصح هامبلتون (1993) بإجراء دراسة عناصر بنية نسخ اختبارية في لغات متعددة كطريقة نافعة للتقييم عما إذا كانت ترجمة الاختبار من لغة المصدر إلى اللغة المستهدفة جيدة إلى حد كافٍ.

7- يجب على المطورين/ الناشرين توفير دليل إحصائي على تكافؤ الأسئلة لكل المجموعات المقصودة. إن الوظيفة التفاوتية للبند وتحليل تخير البند يمكن استخدامها للتقييم إذا كانت وظيفة البند متساوية عبر المجموعات بعد أخذ عامل الاستطاعة بالحسبان. جرى تقديم هذه الطرق في الفصل الرابع.

تفسير الدرجات واستخدام الاختبار

يمكن اعتبار عملية اختبار الإنجاز والأهلية وسيلة لوصول هدف. يقدم الاختبار درجات يفسرها مستخدمون مختلفون لأغراض متنوعة. في مواقف الاختبار التربوي، يجري الظن أن الدرجات لها مدلول عن إمكانية الطلاب أو عن براعتهم في حقل معرفة خاص أو عن معلوماتهم. في دراسات عبر الثقافات، يوفر الاختبار أساساً للقيام بمقارنة بين لغات ومجموعات ثقافية مختلفة وبذلك يكون هناك إدراك أفضل للاختلافات والتماثل بين تلك المجموعات (هامبلتون وبورلورك، 1991). على كل حال فإن تفسير نتائج اختبارات الإنجاز والأهلية التي تجريها مجموعات يتكلمون لغات مختلفة ليست مهمة واضحة لمطورين ومستخدمين الاختبار لأن وجود اختلافات غير مقصودة في صعوبة ومحتويات الاختبار فقط تساهم في إحداث اختلاف الدرجات بين المجموعات أو الأفراد (سيرغي، 1997).

تصرف مطورو الاختبار في السابق وكأن العامل المهم الوحيد في تكيف اختبار لثقافة مختلفة هو ترجمة اللغة الأصلية المستخدمة في الاختبار إلى لغة جديدة. كما ذكر سابقاً في قسم آخر من هذا الفصل فإن ترجمة اختبار من لغة إلى أخرى لا يضمن المساواة في الدرجات عبر اللغات المستخدمة أو الثقافات (انجوف وكوك، 1997، كيسنجر، 1994، هامبلتون، 1993، بريتو، 1992، سيرغي، 1997).

وكما صرح هامبلتون (1994) فإن استخدام طريقة غير جدية تعالج تكييف الاختبار تقود إلى مفهوم خاطئ وهو أن اختلافات الدرجات بين نماذج أو مجموعات يمكن تفسيرها وكأنها حقيقة.

جرت الإشارة إلى أن ترجمة اللغة هو عامل واحد، بالرغم من أنه عامل مهم في تكييف الاختبار، ليس كافياً لوحده لجعل نتائج نسخ اختبار مختلفة اللغات قابلة للمقارنة. إن اختلافات خاصة في أعراف لغوية تمشي يداً بيد مع اختلافات محددة

في الأفكار والسلوك، وهذا يجعل من المستحيل فصل اللغة عن الثقافة. بذلك يثبت أن اعتبار اللغة فقط ليس كافياً.

بالإضافة إلى عامل اللغة، هنالك عدة عوامل يجب أن تؤخذ بعين الاعتبار إذا أردنا أن تكون درجات الاختبار الذي جرى تكييفه للاستخدام في لغات وثقافات متعددة ذات مدلول في التفسير. تتضمن العوامل التي تؤثر على استطاعة مُستخدم الاختبار لاستنتاج تفسير جيد: شروط إدارة الاختبار، المنهاج، السياسة التربوية، دوافع الممتحنين، الحالة الاقتصادية، مستوى الحياة، القيم الثقافية، بنية بنود الاختبار غير المألوفة، قلق من الاختبار، وسرعة الاختبار (هامبلتون وكانجي، 1993؛ فان دي فيفر وبورتينفا، 1991).

وصف هامبلتون وكانجي (1993) بالتفصيل بعض العوامل المهمة التي يجب أن تؤخذ بعين الاعتبار عند تفسير نتائج اختبار الإنجاز في دراسات عبر الثقافات، التماثل في المناهج هو أحد العوامل الذي يجب الانتباه إليه بجدية. إن أي مقارنة على مستوى الإنجاز بين ثقافتين مختلفتين ستكون ضعيفة إلا إذا أخذت الاختلافات في المناهج بالحسبان. هنالك الكثير من الأمثلة عن تأثير اختلافات المناهج في المطبوعات. على سبيل المثال، في النظرة الأولى يبدو أن نتائج الدراسة الثانية للرياضيات العالمية (SIMs) تشير إلى أن أداء طلاب الولايات في بعض المستويات أقل من نظرائهم في اليابان في كل درجة أداء وكل أوجه الرياضيات، على كل عندما لوحظت الفروق في المناهج وضبطت لم تعد هناك أي فروق واضحة بين أداء طلاب الولايات المتحدة والطلاب في اليابان (وستبري، 1992).

إن تأثير حوافز الطلاب على درجات الاختبار شيء يجب التنبه له عند تفسير درجات النتائج في كل الاختبارات: على كل، إن موضوع الاختلافات الثقافية تعقد تأثير ذلك العامل بشكل أبعد. ألقى هامبلتون وكانجي (1993) الضوء على نتائج وينر (1993) التي تساءل فيه عما إذا كان يمكن فصل الخبرة الظاهرة المقاسة بالاختبار. بأي حال أشار وينر إلى دراسة تقويم التقدم التربوي العالمي (IAEP)

(لابوينت، ميد، واسكو، 1992) كدليل على ذلك. في تلك الدراسة كان أداء الطلاب الكوريين أعلى بكثير من نظرائهم الأميركيين، على كل تم إخبار الطلاب الكوريين بأنه تم اختيارهم للمشاركة في تلك الدراسة، كان ذلك شرفاً كبيراً لهم لمدارسهم ولبلدهم، وأن عليهم الأداء بأفضل ما يمكنهم. بالجانب الآخر لم يعط الطلاب الأميركيون أي تلميح لتحفيزهم؛ لذلك شاركوا في اختبار تلك الدراسة وكأنها نشاط مدرسي آخر.

إن فهم تأثير العالم السياسي/ الاجتماعي وجه مهم آخر في تفسير الدرجات لذلك يكون القيام بمقارنة الدرجات بين بلدان متطورة وبلدان غير متطورة ليس بالعملية السهلة الدقيقة. إنها تتطلب فهم المصادر المتوفرة والنوعيات المختلفة للخدمات التربوية التي يمكن أن تؤثر على أي قرارات عن القدرة الحقيقة التي تعكسها درجات الاختبار (هامبلتون وكانجي، 1993، اوليدو، 1981).

وضع هامبلتون وفان دي فيفر (1996) مع اللجنة الدولية للعلماء النفسيين وعلماء القياس السيكولوجي مجموعة من الخطوط العريضة، لتكييف اختبارات تربوية ونفسية. كان ضمنها بعض خطوط التوثيق وتفسير الدرجات المفيدة جداً للممارسين الذين يحاولون الاستفادة من تلك الدرجات.

فيما يلي الخطوط العريضة للتوثيق وتفسير الدرجات:

1- عندما يجري تكييف/ ترجمة الأداة للاستخدام في مجموعة أخرى، يجب توثيق التغييرات مع دليل التكافؤ. أصر فان دي فيفرو هامبلتون على أن فهم أي تغيرات جرت لتعزيز صدق الأداة المكيفة مهم لمستخدمي الاختبار عند تقرير عما إذا كانت أداة ما مناسبة لأغراضهم في البيئة الجديدة.

بالإضافة إلى المعلومات عن التغييرات التي جرت على الأداة، يجب أن يكون لدى المستخدمين إمكانية للحصول على معلومات عن تكافؤ نسخ الاختبار في لغة المصدر واللغة المستهدفة، مواصفات عملية الترجمة ونتائج تحليل تحيز البند أو تحليل العوامل. من المهم أيضاً لمستخدمين الاختبار معرفة عما إذا تم أخذ عوامل ثقافية معينة بالاعتبار عند وضع الاختبار.

2- يجب أن لا تُقوم الدرجات المختلفة للمجموعات التي أجرت اختبار الأداة حسب القيمة الظاهرية. تقوم على البحث مسؤولية إثبات الاختلافات بدلائل تجريبية. من الضروري الانتباه أن معنى الاختلافات بين مجموعات مختلفة يمكن تفسيره بطرق مختلفة. نبه فان دي فيفر وهامبلتون إلى أنه في حال عدم اختيار الباحث تقبل تفسير معين للدرجات، فيجب عليه توفير الدليل لدعم ذلك الخيار.

غالباً تتطلب مجموعة الدلائل مقاييس العوامل مختلفة الخيارات. إن الاختبار الذي جرى تكيفه حسب عملية تقنية موثوقة تتطلب جهداً أقل في دعم تفسير الدرجات لأن صدق الأداة جرى إثباتها إلى حد ما. أشار فان دي فيفر وهامبلتون إلى أنه حتى في أحسن الظروف فإن على الباحثين القيام بكل الجهود لإظهار تفسير دقيق لنتائج نسخ اختبار متعددة.

3- يمكن إقامة مقارنة للمجموعات المختلفة على المستوى الثابت الذي جرى إقامته للمقياس الذي بينته الدرجات. أشار فان دي فيفر وهامبلتون هنا إلى مفهوم قياس الدرجات. جرت مناقشة موضوعات متعلقة بإقامة تكافؤ قياس المقياس في هذا الفصل من قبل. من الممكن عند توفر عدد كبير من النماذج إقامة درجات نسخ اختبار للغات مختلفة على مقياس واحد لإجراء مقارنات للمفهوم. إن النقطة الأساسية لتلك الخطوط العريضة هي حث الباحثين للمرة الثانية على عدم القيام بمقارنة درجات نسخ مختلفة من الاختبار لا مبرر لها إلا إذا توفر دليل الصدق.

4- يجب على المطورين/ الناشرين توفير معلومات دقيقة عن الطرق التي يمكن أن تؤثر فيها الثقافة الاجتماعية والظروف البيئة للمجموعات على الأداء في الاختبار، كما يجب عليهم اقتراح إجراءات لبيان تلك التأثيرات على تفسير النتائج. شكل فان دي فيفر وهامبلتون في هذا الدليل من الخطوط العريضة طريقة عملية للتعامل مع العامل الثقافية/ الاجتماعية والظروف البيئية التي لها

دور في تفسير الدرجات. إن أفضل طريقة لنقل معلومات مناسبة لمستخدم الاختبار هي توفير كتيب للاختبار الذي يفصل كل المتغيرات التي بحثت عند تطوير الأداة (الصفات الثقافية)، الحالة الاجتماعية/ الاقتصادية، العمر، الجنس، التعليم للمجموعة المستهدفة.

إذا كان مثل تلك التحليلات متوفرة لمستخدمي الاختبار، فسيكون لديهم معلومات أكثر عن كيفية بيان أسباب تلك العوامل عند تفسير الدرجات (براكن وبارونا، 1991؛ فان دي فيفر وبورتينفا، 1991).

لكي تكون تفسيرات الدرجات ذات معنى، من الضروري إثبات أن القياسات المكيفة تُقوّم ذات البنية في اللغة أو الثقافة. جرى مناقشة موضوعات ذات علاقة بتقويم تكافؤ البنية أيضاً في أوائل هذا الفصل.

إن التحديد على أنه يجب على القياس المكيف تغطية ذات الأبعاد بذات النسب في اللغات والثقافات المختلفة للمجموعات ضروري جداً لعملية تفسير جيدة ولاستخدام درجات الاختبار (أيسينك وأيسينك، 1983). كما ذكر سابقاً، هناك تقنية واحدة لإقامة تحليل هذا العامل.

بإمكان معلومات معيارية أيضاً توفير معلومات مهمة لشخص محترف يحاول استخراج معنى من الدرجات وذلك بوضع شخص ما ضمن مجموعة من الذين يقوم بإجراء الاختبار. على كل حال من الضروري جداً إثبات أن المعايير المطورة لاختبار تم إجراؤه لمجموعة لغوية وثقافية مناسبة لتفسير درجات مجموعة لغوية وثقافية مختلفة (انظر كيسنجر، 1994، مناقشة موضوعات متعلقة باستخدام معلومات عن معايير للتقويم عبر الثقافات/ اللغات).

إدارة الاختبار

كما ذكر سابقاً في هذا القسم من الفصل، إن تحيز الطريقة مصطلح عام يمكن أن يشير إلى أي عامل تهديد للصدق المرتبطة بظروف إدارة الاختبار (فان دي فيفر وهامبلتون، 1996). إن عدم الخبرة في بنية البند، في هيكل الاختبار، أو في

وضع الاختبار، كل هذا قد يؤدي بشكل عام إلى ذلك التحيز. هناك أوجه أخرى للإدارة مثل، حالة الغرفة، حافز الاختبار، التأثيرات الإدارية، ومشكلات في التواصل بين الإداري والشخص الذي يأخذ الاختبار يمكن أن تؤدي إلى تحيز في نتائج الاختبار. قد يوجد تحيز المنهج إلى حد ما في كل مقارنات نتائج اختبار الإنجاز والأهلية عبر الثقافات ويمكن أن يقود إلى تفسير الاختلافات في النتائج بين المجموعات التي سببتها إجراءات الاختبارات وكأنها تفسير اختلافات حقيقية في المقدرة (هامبلتون وكانجي، 1993؛ فان دي فيفر وهامبلتون، 1996).

يمكن دراسة التحيز الحاصل بسبب الإجراءات الإدارية بطرق عديدة. إحدى تلك الطرق هي إعادة إدارة الاختبار. إن دراسة تغيرات تماثل الدرجات عبر الثقافات في إدارة الاختبار الأول ثم إدارة الاختبار الثاني يمكن أن تعطي دلائل مهمة عن صدق الاستنتاجات التي تم الحصول عليها.

عندما يحصل طلاب من مجموعات مختلفة على درجات اختبار متساوية في الاختبار الأول ودرجات مختلفة في الثاني، يمكن أن يكون هناك تساؤل عن صدق استنتاج الدرجات التي تم الحصول عليها من الاختبار الأول.

هناك طريقة ثانية لدراسة التحيز بسبب طريقة إدارة الأداة وهي استخدام الأداة طريقة غير قياسية. يكون ذلك ممكناً عند القيام باستخلاص كل المعلومات والإجابات من الطلاب عن تفسير الإرشادات، البنود، إجابة الخيارات، والحوافز عند اختيار أسئلة محددة (فان دي فيفر وهامبلتون، 1996).

هنالك الكثير من الطرق لتجنب مشكلات متعلقة بتحيز الإدارة. إحدى الطرق التي ينصح بها هي التأكد من أن إرشادات الاختبار نفسه واضحة وتُسيّر نفسها بنفسها، مع أقل اعتماد على التواصل الشفهي مع الإداري (فان دي فيفر وبورتينغا، 1991). في كل مواقف الاختبار من الضروري فهم الثقافات التي يأتي منها الطلاب الذين يجرون الاختبار ويجب أن يكون الإداريون قادرين على التواصل في لغة هؤلاء الطلاب على الأقل (كيسنجر، 1994؛ كيسنجر وكارلسون، 1992).

خلاصة لذلك من المفضل أن يكون إداريو الاختبار من قومية المجموعة الآخذة للاختبار ذاتها، أو أن يكون لهم معرفة بالثقافة واللغة، وأن يتمتعوا بالخبرة والمعرفة بالإدارة وأن يملكوا خبرة في القياسات (هامبلتون، كانجي، 1993). من الضروري أيضاً تشجيع ثبات إجراءات إدارة الاختبار عبر المجموعات المختلفة التي يتم اختبارها. إن أفضل طريقة لذلك هو توفير تدريب مستمر لجميع إداريي الاختبار. يجب أن يؤكد ذلك التدريب على وضوح وعدم غموض عملية التواصل بين الإداريين وآخذي الاختبار، على أهمية اتباع إرشادات الاختبار بدقة، التقيد بالوقت المحدد وعلى التأثير المحتمل لإداري الاختبار على صدق الاستنتاجات التي تم الحصول عليها من الدرجات (هامبلتون وكانجي، 1993).

قدم فان دي فيفر وهامبلتون (1996) خطوطاً عريضة واضحة ودقيقة لإداريي الاختبارات المكيفة:

1- يجب على مطوري الأداة والإداريين محاولة توقع كل أنواع المشكلات التي يمكن حدوثها واتخاذ الإجراءات المناسبة لمعالجة تلك المشكلات وذلك بإعداد مواد وإرشادات مناسبة. كما أشار فان دي فيفر وهامبلتون فإن توقع مشكلات إدارية ليس معقداً جداً. يمكن جعل تلك المهمة سهلة وذلك عند إقامة دراسة استطلاعية تستخدم الاختبار بطريقة غير قياسية لاستخلاص إجابات مختلفة من الطلاب، يمكن أن تساعد الملاحظة الدقيقة والتغذية الراجعة من الطلاب في الكشف عن التأثيرات الإدارية المحتملة.

2- يجب على إداريي الأداة إدراك بعض العوامل المتعلقة بالمواد المحفزة، الإجراءات الإدارية، طرق الإجابة التي يمكن أن تُقوم صدق الاستنتاج التي تم أخذها من الدرجات. بالرغم من أن الترجمة الحرفية كانت تُفضل دائماً فإن على الإداريين معرفة المشكلات التي يمكن أن تحدثها تلك الترجمات. على سبيل المثال يمكن أن تكون هنالك بعض الأوجه الضمنية في الإرشادات لم يجر إيصالها من خلال الترجمة.

3- يجب أن تكون أوجه البيئة المحيطة التي تؤثر على إدارة الأداة متشابهة إلى أبعد حد عبر المجموعات التي تستهدفها تلك الأداة. من المعروف أن السيطرة على وضع البيئة في البحث الميداني مستحيل. لكن على الإداريين أن يدركوا العوامل البيئية المختلفة التي يمكن أن تؤثر على صدق الدرجات لكي يقوموا بكل جهد لجعلها ثابتة.

4- يجب أن تكون إرشادات إدارة الأداة موجودة في اللغتين: لغة المصدر واللغة المستهدفة وذلك للتقليل من التأثير غير المرغوب به للاختلافات. بالإضافة إلى ذلك فإن إرشادات اختبار مطولة تحتوي على تمارين وأمثلة مختلفة تساعد على تقليل الاختلافات المتعلقة بالإدارة نفسها.

5- يجب أن يحدد كتيب الأداة كل أوجه الأداة وإدارتها التي تتطلب التدقيق في استخدام الأداة في مفهوم ثقافي جديد، أشار فان دي فيفر وهامبلتون أنه بما أن مطوري الاختبار يعملون على تكييف الاختبار للاستخدام في ثقافات ولغات مختلفة فإنهم سيكشفون عن موضوعات محددة عن استخدام ذلك الاختبار في المحيط الثقافي الجديد.

يستفيد إداريو الاختبار بمعرفة تجربة مطوري الاختبار ويجب أن يدركوا المشكلات المحتملة لكي يتجنبوا إعادتها.

6- يجب أن لا يكون الإداريون فضوليين ويجب الإقلال عن التفاعل بين الإداريين والذين يأخذون الاختبار. يجب اتباع القواعد المذكورة في كتيب الأداة. تنتج أخطاء شائعة ومهمة في مقارنة درجات عبر الثقافات بسبب التفاعل غير المحكم بين الإداريين والطلاب. يجب أن يحدد كتيب إداريي الاختبار تلك المشكلات الحالة ويقدم الحلول المناسبة.

الخلاصة

كان الهدف من هذا الفصل مراجعة الموضوعات المنهجية المتعلقة بتكييف قياسات الإنجاز والأهلية. ركزنا في هذا الفصل على ست نواح:

(1) تكافؤ البنية.

(2) تكافؤ وحدات القياس.

(3) ترجة الاختبارات ومواد الاختبار.

(4) متغيرات البند الوظيفية.

(5) تفسير الدرجات.

(6) إدارة الاختبار.

وجدنا أن التقدم الحاصل في تلك النواحي في العقد الماضي مطمئنة جداً بحيث أصبح الباحثون المهتمون بالمقارنة عبر اللغات/ الثقافات لقياسات الإنجاز والأهلية أكثر إدراكاً لتعقيدات الموضوعات التي تؤثر على القدرة للقيام بمقارنة تقويم درجات صادقة.

نعزو نمو ذلك التطور إلى عوامل عدة. واحد منها هو أن تعدد القوى السياسية والاقتصادية التي تؤدي بدورها إلى الاقتصاد العالمي مع تدفق الهجرة جعل من الضروري إجراء الاختبار في لغات عديدة. كنتيجة لذلك أصبح كثير من الممارسين يهتمون بتلك الناحية للتقويم وأصبحوا يثيرون انتباه الباحثين إلى مشكلات عملية متنوعة. بالإضافة إلى ذلك أصبح كثير من المنهجيين مثل IRT وبعض نماذج تعادل البنية يهتمون بتلك الأمور.

نتنبأ بأن العقد التالي سوف يظهر تضاعفاً للحاجة ولاستخدام الاختبارات المكيفة. مع الاستخدام المتوسع للتقويم الذي خصص لثقافة ولغة واحدة والذي جرى استخدامه لمجموعات ذات ثقافات ولغات مختلفة سيصبح هناك معرفة متضاعفة عن المهارات والأهلية المقاسة عبر المجموعات الثقافية بالإضافة إلى طرق محسنة للقيام بالمقارنات الضرورية.

شكر

يرغب الكتَّاب بالتعبير عن شكرهم لمشاركة دانيل أنجور بالعمل في هذا الفصل. له التقدير الكبير لمساعدته في التحرير.

*****

المراجع

Angoff, W.H., & Cook, L. L. (1988). *Equating the scores of the "Prueba de Aptitud Academica" and the "Scholastic Aptitude Test"* (Report No. 88-2). New York: College Entrance Examination Board.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185–216.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29,* 543–553.

Ercikan, K. (1999, April). *Translation DIF in TIMSS.* Paper presented at the meeting of the National Council of Measurement in Education, Montreal, Canada.

Ercikan, K. (2000). Disentangling sources of differential item functioning in multilingual assessments. *International Journal of Testing, 2,* 199–215.

Everson, H. T., Guerrero, A., & Laitusis, V. (1998, April). *Preliminary evidence of construct equivalence of mathematics tests administered across languages: An analysis of findings from the SAT I and the Prueba de Aptitud Academica tests.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Eysenck, H. J., & Eysenck, S. B. G. (1983). Recent advances in the cross-cultural study of personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 41–69). Hillsdale, NJ: Lawrence Erlbaum Associates.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Geisinger, K. F., & Carlson, J. F. (1992). *Assessing language-minority students* (Report No. EDO-TM-92-4). Washington DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Gierl, M. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*(1), 57–58.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (1996, March). *Guidelines for adapting educational and psychological tests.* Paper presented at the meeting of the National Council of Measurement in Education, New York.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission, 18,* 229–244.

Hambleton, R. K., & Kanjee, A. (1993, April). *Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods.* Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

Hambleton, R. K., & Kanjee, A. (1995). Translating tests and attitude scales. In T. Husen & T. N. Postlewaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 6328–6334). New York: Pergamon.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45,* 153–171.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296–309.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics* (Report No. 22-CAEP-01). Princeton, NJ: Educational Testing Service.

Olmedo, E. E. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Perez, A. (1975). *Measurement of bilingual ability.* Unpublished master's thesis, University of Puerto Rico, San Juan.

Poortinga, Y.H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W Berry (Eds.), *Human assessment and cross-cultural factors?* (pp. 237–258). New York: Plenum.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737–756.

Poortinga, Y. H. (1995). Uses of tests across cultures. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 187–206). Boston: Kluwer Academic.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Reckase, M. D., & Kunce, C. (1999, April). *Translation accuracy of a technical credentialing examination.* Paper presented at the meeting of the National Council of Measurement in Education, Montreal, Canada.

Sireci, S. G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice, 16,* 12–19.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 147–165.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Stanley, J. C., & Hopkins, K. D. (1972). *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice-Hall.
van de Vijver, F. J. R. (2002). Cross-cultural assessment: Value for money? *Applied Psychology: An International Review, 51*(4), 545–566.
van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating test: Some practical guidelines. *European Psychologist, 1*, 89–99.
van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33–51.
van de Vijver, F. J. R., & Lonner, W. (1995). A bibliometric analysis of the *Journal of Cross-Cultural Psychology, 26*, 591–602.
van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht, Netherlands: Kluwer Academic.
van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263–279.
Van der Flier, H., & Drenth, P. J. D. (1980). Fair selection and comparability of test scores. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates*. New York: Wiley.
Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1–21.
Westbury, I. (1992). Comparing American and Japanese achievement: Is the United States really a low achiever? *Educational Researcher, 21*, 18–24.
Woodcock, R. W. (1985). *Woodcock Language Proficiency Battery, Spanish Form: Technical summary* (Assessment Services Bulletin No. 9). Allen, TX: DLM.

القسم الثاني

تكييف التقاطع اللغوي للاختبارات النفسية والتربوية: تطبيقات على اختبارات تتعلق بالإنجاز والأهلية والشخصية.

8

تكييف الاختبار في برنامج مصدّق (معتمد) واسع النطاق

سيندي ت. فيتزجيرالد

التصديق أو الاعتماد (Certification) هو نشاط طوعي مفصل يقوم بسد الهوة بين الاعتمادات الأكاديمية وبين الطلب المتزايد والحاجة الملحّة لتطور سريع في مكان العمل. لقد تطورت برامج الشهادات الاحترافية بخطوات سريعة في السنوات العدة المنصرمة. وقد أشار مك كيليب وكوكس (1998) أن هناك ما يزيد عن 700 برنامج مصدّق في الولايات المتحدة فقط. بتوفر الحاسوب الذي يصدر الاختبارات فإن إمكانية هذه البرامج لإصدار الاختبارات إلى السوق لكل أنحاء العالم بات الآن شيئاً ممكناً (مثلاً: ميلس، بوتينزا، فريمر ووارت، 2002) في المعلومات التقنية في مجال (IT) تقوم عدة شركات مثل مايكروسوفت ونوفل بتسليم الاختبارات التي تعتمد على الحاسوب إلى أكثر من 100 بلد. إلا أن إصدار اختبار ما باللغة الإنكليزية ليس كافياً من أجل تلبية الطلب العالمي لبرامجها المتعلقة بالشهادات فإن على مايكروسوفت والشركات الأخرى أن تكيف فحوصاتها إلى لغات وثقافات متعددة.

إن التحرير والثقافة والمصطلحات الفنية والتغيرات في نوعية الأفراد الذين يقومون بالترجمة أو بالتنقيح كفيلة بأن تصنع مشكلة كبيرة لوكالات الاختبار. على سبيل المثال تختلف اللغة اليابانية عن اللغة الإنكليزية أو الألمانية. إن صياغة فحوصات مصدّقة في مجال الـ IT ليست كبناء اختبار إنجاز أو الاستعداد لاختبار.

إن التحول سريع جداً، بحيث لا ينبغي فقط أن ينفذ بشكل صحيح ولكنه بشكل سريع أيضاً. لقد تمت الإشارة إلى عملية تكييف الاختبارات إلى لغات عدة من قبل عدة شركات مركزية الـ IT. إن العديد من الشركات ذات المقاييس ذاتها التي تم إخضاعها لمركزية منتجات برامج الحاسوب بحاجة إلى أن تطبق عليها فحوصات مصدّقة أو تمت مركزتها*.

هناك العديد من العوائق التي يجب التغلب عليها في التطويرات المتعلقة بالاختبار المرتكز على الحاسوب (انظر إلى ميلس وآل، 2002) تتزايد هذه العوائق بعشر مرات حين يتم تطوير هذه الاختبارات لغرض برنامج مصدّق لـ IT الذي يقوم باختبار مهارة الفرد مع الأنظمة المطبّقة مثل Windows 2000 أو Windows NT Server 4.0 مع اختبار مهارة الفرد لتطور تطبيقات برامج الحاسوب.

هناك أربع عوائق رئيسة يجب التغلب عليها لتطوير الاختبار المصدّق لـ IT أولاً، يجب أن يتم تطوير الاختبارات بشكل سريع، لأن برامج الحاسوب بحاجة إلى التحديث باستمرار، ويجب أيضاً أن يتم تحديث الاختبارات المصدّقة. يجب أن يتم أيضاً تحديث الفحوصات المصدّقة والمناهج التعليمية المطبقة على الحاسوب بنفس السرعة؛ لأن برامج الحاسوب والمناهج التعليمية المطبقة على الحاسوب وتطور الاختبار تمت بشكل مترادف. وقد تم التشديد على المصادر بصرامة على الرغم من أن شركة مايكروسوفت تدرج المساهمات التقنية الخارجية في عملية تطوير الاختبار، فإن مطوري الاختبار لا يزالون يعتمدون بشكل كبير على خبراء موضوعات البحث الداخلية (SMEs)، الذين يقومون بدورهم بتطوير منتجات برامج الحاسوب ذاتها. ثالثاً، تحتاج الاختبارات إلى أن تُمركز إلى لغات متعددة. وهذا يزيد بشكل جوهري المقاييس التي يجب أن تطبق على عملية تطور الاختبار. لقد طورت لجنة الاختبار العالمية 22 (ITC) دليل لتكييف الاختبارات من لغة إلى أخرى (ملخص من

منتدى سور الأزبكية
www.books4all.net

(*) المركزة تعني تحويل شيء لجعله مستخدماً محلياً.

قبل هاملتون، 1994، انظر أيضاً المقطع 1 هذا الكتاب). إن الالتزام بدلائل تكييف الاختبار ITC ومقاييس مايكروسوفت الإضافية تجعل من كتابة وتكييف المادة شيء معقد و مستهلك للوقت. إن الاختبار على الحاسوب يسبب تعقيدات إضافية: إن الاختبارات في كل اللغات بحاجة إلى التنقيح في صيغتهم المرتكزة على الحاسوب لأغراض تتعلق بتأكيد الجودة. على سبيل المثال، هذا التنقيح يتضمن مقارنة صور الشاشة لاختبار ما، مع السطح البيّني الفعلي لمنتج برنامج الحاسوب المركز.

أخيراً، بسبب سرعة تطور الاختبار وبسبب الحاجة إلى مركزة الفحوصات هناك العديد من المسائل المتعلقة بقياس سرعة ودقة العمليات العقلية والصدق التي يجب توجيهها.

يركز هذا المقطع على شرح الطريقة التي يتم فيها التكييف والتصديق على الفحوصات المعتمدة في شركة مايكروسوفت وذلك بهدف الاستخدام في لغات متعددة. إن التوقعات هي أن هذا المقطع سيقوم بإلقاء الضوء على التحديات والطرق التي يمكن أن تواجه بها هذه التحديات من قبل مؤسسات أخرى. تم ترتيب إشعار التذكير لهذا المقطع حول الخطوات الرئيسة المدرجة في تكييف الاختبارات المعتمدة من اللغة الإنكليزية إلى اللغات الأخرى. وقد تم توضيح كل خطوة عبر مثال (انظر أيضا هامبلتون، سيرسي وروبن، 1999) قبل شرح هذه الخطوات، تم عرض بعض الخلفيات التي يرتكز عليها البرنامج ومجالاته.

نظرة عامة على برنامج احترافي مصدق (معتمد) لمايكروسوفت

لقد أصدرت شركة مايكروسوفت ما يزيد عن مليون اختبار كل سنة لأكثر من 75 بلداً. ويتوفر حالياً 42 اختباراً في أنحاء العالم. بالإضافة إلى اللغة الإنكليزية، تم تكييف كل اختبار للإدارة المحلية فيما يقدر بـ 13 لغة. في الوقت الحالي، تقوم شركة مايكروسوفت بتقديم اختبارات مصدقة في اللغات التالية: الإنكليزية واليابانية والكورية والصينية المبسطة والألمانية والهنجارية والبولندية والفرنسية والروسية والإيطالية والإسبانية والبرتغالية البرازيلية والتشيكية.

تاريخياً، تم تكييف اختبارات الشخصية واختبارات الـ IQ أما الآن فهناك نقلة نوعية للإصدار العالمي للفحوصات في لغات أخرى من أجل المهن، كالتقنية المعلوماتية، والطب والأمن والمحاسبة. بل حتى في اختبارات اختبار القبول في الثانوية مثل اختبار التأهيل المدرسي المتوفر الآن في اللغة الإسبانية (مثلاً، انظر إلى المقطع 7، هذا الكتاب).

مميزات اختبار مصدق (معتمد) لشركة مايكروسوفت

تم تطوير اختبارات مايكروسوفت المصدقة بإدخال محترفين تقنيين في الصناعة، وهذا يعكس كيف تستخدم منتجات مايكروسوفت في المؤسسات عبر العالم. تحوي اختبارات مايكروسوفت المصدقة، إجمالاً، أنماط الأسئلة التالية:

- أسئلة تقليدية متعددة الخيارات (MCQ) تقوم بقياس المعرفة الأساسية وفهم لمنتجات مايكروسوفت والتقنيات.
- أسئلة ترتكز على السيناريو تقيس قدرة المرشحين على تحليل المواقف.
- أسئلة تقديرية متعددة ترتكز على السيناريو تقيس قدرة المرشحين على تحليل وتركيب المعلومات ثم تقييم جودة حل مقترح.
- أسئلة صورية تقيس قدرة المرشحين على استخدام النسخة المزيفة لمنتج برنامج الحاسوب. وهي تقويمات حقيقية موثوقة لقدرة الممتحنين على استخدام منتج برنامج الحاسوب لإكمال المهام المحددة.
- أسئلة الإشارة والنقر التي تقيس قدرة المرشحين على تحديد مكان في صورة بيانية ما. مثال على أحد هذه الأسئلة، قد يطلب من الممتحنين أن يضعوا المشيرة (الفأرة) على جزء من مخطط الشبكة الذي يطابق (Server) المضيف لموقع الشركة على شبكة الإنترنت.
- أسئلة ترتكز على الجر والحذف تقيس قدرة المرشحين على ترتيب المعلومات (نص أو صورة بيانية) وذلك بتحريكها من شاشة لوضعها بأخرى. مثال على ذلك قد يطلب من الممتحنين أن يقوموا بتصميم شبكة. لتحقيق ذلك،

سيقوم الممتحن بسحب ثلاثة إيقونات لمحطة عمل إلى لوحة ملونة ثم ربطها بمخدم ما.

بالإضافة إلى هذه الأنماط من الأسئلة، فإن شركة مايكروسوفت تقوم باستعمال حالة حروف معقدة للسيناريوهات الدراسية وتعتمد بشكل كبير على استعمال الصور البيانية والجداول ومعروضات أخرى. لمعلومات إضافية حول أنماط النظام المحددة هذه والجوانب الأخرى لبرامجهم المصدّقة أو المعتمدة، انظر إلى صفحاتهم على الشبكة في http://www.microsoft.com/learning/mcp/، وإلى أطروحة العرض الممتازة التي تصف أنماط النظام للأسئلة البارزة المعدة من قبل زينسكي وسيرسي (2002).

تكييف اختبارات اللغة الإنكليزية للاستخدام العالمي

تتألف عملية تكييف الاختبار في شركة مايكروسوفت من أربع مراحل: مرحلة تطوير الاختبار باللغة الإنكليزية والمرحلة المركزة والمرحلة ما قبل المركزة.

إن التطور في صيغة اللغة الإنكليزية للاختبار يتبع خطوات تطور الاختبار التقليدي لكل من إدارة تحليل مهمة عمل ما، وتطور السؤال وأسئلة اختبار الاختصاص، وتحليل السؤال، وتركيب الصيغ والإعدادات النموذجية.

أما مرحلة ما قبل المركزة، فتتم بشكل متزامن مع المراحل الأخيرة لتطوير اختبار اللغة الإنكليزية. يقوم المترجمون خلال مرحلة ما قبل المركزة باستعراض أسئلة الاختبار الإنكليزية وذلك من أجل التنبؤ بالمشكلات التي يمكن أن تنشأ من العمل بمنتجات ممركزة.

تتألف مرحلة المركزة من ترجمة مضمون الاختبار المرتكز على ملفات اختبار اللغة الإنكليزية النهائي. يتم تزويد المترجمين بالتدريب وبالدلائل لإكمال التراجم. يتم إعطاء تعليمات إلى المترجمين لترجمة مضمون السؤال بدلاً من الترجمة التي تعتمد على ترجمة كلمة بكلمة.

المرحلة النهائية وهي مرحلة ما بعد المركزة، تدرج عرض تقني مكثّف وإصدار للاختبار. ينفذ العرض التقني من قبل SME في البلد الأصلي. يتم تزويد المنقحين بنسخة إلكترونية للاختبار بحيث يستطيعون رؤيته تماماً كما سيظهر للممتحنين. يقوم المنقحون بتقديم تغذية استرجاعية مباشرة إلى نسخة الاستعراض الإلكترونية لهذا الاختبار حالما يتم تأكيد التغذية الاسترجاعية، فإنه يتم إعادة تحرير وإعادة نشر الاختبار ثم يصدر إلى مختلف أنواع العالم. قد يخضع المتحنون للاختبار في اللغة التي يختارونها. إن إشعار تذكير هذا المقطع يدور حول تفاصيل كاملة عن كل من هذه المراحل، ويقدم أمثلة على الأدوات والصيغ وقائمة المراجعة إلى آخره، التي يتم استعماله.

تطوير اختبار اللغة الإنكليزية

إن تطوير اختبار اللغة الإنكليزية هو شيء نموذجي للغاية من تحليل مهمة العمل وتركيب مواصفات الاختبار إلى أسئلة اختبار اختبار الاختصاص، إلى الإصدار الجديد للاختبار. إن الشيء الفريد هو أن هذه الأسئلة ربما قد تم تكييفها إلى لغات أخرى، فقط أعطي الاهتمام الأكبر لاختيار المضمون بحيث تم تعميمه عبر اللغات والثقافات. مثال عن ذلك، إذا كان جزء من منتج برنامج الحاسوب ليس متوفراً في المنتج المركز (قد يكون هذا وظيفة أو ميزة تتعلق بالمنتج)، عندها لا يمكن أن يكون جزء أساسي بالتقييم. علاوة على ذلك، فقد تم اختيار القضايا المدروسة مع الأخذ بعين الاعتبار كيف ستعمل بصيغة مكيّفة. تشمل الأمثلة تطوير تطبيق مرتكز على كل من القوانين الدولية والأمراض الأكثر انتشاراً والرياضيات ومبادئ المحاسبة وما شابه. تتبع هذه العملية ما يشار إليه عادةً في الأدب باللا مركزية. باختصار فإن شركة مايكروسوفت تحاول أن تتنبأ بالمشكلات وأن تقدم أمثلة وثيقة الصلة عالمياً.

ما قبل المركزة

على الرغم من أن محرري الأسئلة ومطوري الاختبار حساسون للمسائل المتعلقة باستخدام هذه الاختبارات في لغات متعددة، فإن هذه العملية هي على الأغلب غير رسمية. إن خطوة ما قبل المركزة هي محاولة لجعل هذه الخطوة أكثر رسمية. وهذا يتضمن توظيف المؤسسات الصغيرة والمتوسطة التي تقوم بلفت انتباه فريق تطوير الاختبار إلى أية اختلافات في المنتج نفسه أو إلى الطريقة التي يستعمل بها في البلدان الأخرى. تعد هذه المرحلة حاسمة لأن وظيفة برامج مايكروسوفت ليست دائماً متطابقة عبر اللغات. ويعزى ذلك إلى العوائق مثل توفر مكونات صلبة معينة خاصة بالحاسوب. أيضاً أعيد النظر بالأسئلة لتحديد ما إذا كانت تلبي عدد من المقاييس الإضافية كالقدرة على جعل السيناريوهات محلية، أسماء المُخدّم (Server) والصور البيانية وعدة مسميات. متوسطياً فإن مرحلة ما قبل المركزة تستغرق حوالي أسبوعين.

إن كل الموضوعات الموجودة خلال هذه المرحلة قد رصدت باستعمال صيغة تحليل المركزة. يقدم جدول رقم 8-1 مثال منجز عن هذه الصيغة. نرى في المثال، تحت عمود المسألة، إن الشخص الذي أكمل الصيغة قد أشار إلى أن هناك سبع مسائل محددة. يحدد الجدول أيضاً.

المصادر التي عرّفت المسائل. ويحدد الطريقة التي سيتم بها حل المسألة وما هي اللغات الأكثر تأثيراً. في المثال المقدم بالجدول رقم 8-1، تم تعريف كل من وحدات القياس والكميات كمسألة (مسألة رقم 1) من قبل الفريق المُحرر ومن قبل مجموعة تأكيد النوعية العالمية (IQA) في شركة مايكروسوفت. إن الحل هو أن على المترجمين أن يستخدموا المقاييس الأكثر شيوعاً في الثقافة. إذاً سوف تحتاج كلمة Units الإنكليزية إلى أن تكيف لتصبح Metric من أجل البلدان الأوروبية. أما المسألة الثانية المعرّفة فيتضمن استعمال تعبير you في اللغات كاللغة الإسبانية

الجدول 1.8

صيغة حلول مركزة (مقتطف فقط)

رقم المسألة	المسألة	المصدر	الحل	اللغة
1	قياس المسافة و/أو النوعية تختلف بين البلدان.	تحرير، مجموعة تأمين النوعية العالمية (IQA)	يمكن للمترجمين استعمال مقاييس تكون أكثر شيوعاً للثقافة	الكل
2	صعوبة بترجمة أنت "you"	تحرير IQA	يمكن للمترجمين أن يستعملوا صيغاً لـ you تكون أكثر شيوعاً للثقافة.	إسباني
3	إن السيناريو في واحد من الأسئلة يتطلب فحصاً صحياً. هل لدى كل البلدان فحص صحي؟	مدير المركزة	ليس لدى كل البلدان فحص صحي. سيتوجب استعمال سيناريو بديل.	هذا صحيح في عدة بلدان خارج الولايات المتحدة
4	إن ترجمة المصطلحات/الأسماء التقنية المتراكمة هو شيء صعب ومرعب تجاه المرشح.	IQA	كل بريد إلكتروني إلى الشركة المترجمة يجب أن يشير المترجمون إلى بائع الـ KIT، مثل كيف اقترحت الـ IQA معالجة التراكم	الكل
5	الصفات المكدسة تجعل التمركز شيئاً صعباً.	كاتب المادة تحرير	نحن مدركون لهذه الصعوبة. يحتاج المترجمون لتقديم هذه الصفات إلى MS لإصدار القرار. إن أملنا أيضاً هو أننا نستطيع تحديد وحل هذا الأمر في تنقيح أعيد تمركزه.	الكل
6	المترجمون غير مطلعين على الألفاظ الأولية لـ MS	IQA	قائمة من الألفاظ الأولية ستسلم في بداية كل مشروع	الكل
7	التزويد بمعلومات تطويق الشفرة. أين يمكن أن تفك؟	مدير البرنامج، كاتب المادة	نحن نعمل في الطرقة الأمثل لمعالجة ذلك. سيكون لدينا دليل عام ولكن لنكن مرنين.	الكل

مثلاً، هناك أكثر من كلمة واحدة يمكن استخدامها لقول (you أنت). إن الحل بالنسبة للمترجم هو أنه يجب استعمال صيغة you الأكثر شيوعاً في الثقافة.

المركزة (التحويل إلى المحليّة)

حالما تكتمل المرحلة المركزة يتم إجراء الترجمة الفعلية لمضمون الاختبار. يحتوي الجدول رقم (8-2) في بند صيغة التنقيح، مجموعة من المعايير لتقييم الأسئلة. إن المعايير في هذا الجدول لا تطبق لتعديل الصيغ اللغوية للاختبار، بل تطبق على اختبار اللغة الإنكليزية. يزودنا هذا الجدول بإطار عمل لتقييم الأسئلة في اللغات المتعددة. تم استخدام الجدول كأداة لأنه لا يوجد مصدر داخلي ينصح المترجمين أين يمكن إجراء التغيير. يقدم لنا الجدول (8-2) مقتطفاً كاملاً لصيغة تنقيح أسئلة ما. رتبت صيغة التنقيح هذه في حوالي ستة أقسام. يضم الأول تعليقات عامة حول السؤال، ثم تعليقات حول السيناريو في حالات تم استعمال السيناريو بها. يتعلق القسم الثالث والرابع بجذر كلمة السؤال وخيارات الإجابة. أما القسم الخامس فيتضمن مواصفات عن الرسوم البيانية والجداول. بينما يقدم القسم السادس بيانات عن صور الشاشة التي يحتمل أو لا يحتمل وجودها.

من الضروري أن يحدد مصدر أو أكثر لكل من التعليقات والمواصفات أن وحدة السؤال قد طابقت المواصفات المقدمة للاختبار. باعتبار أن المرء يستطيع أن يرى من خلال المقتطف المقدم إن كاتب السؤال مسؤولاً عن الضمانة لأن وحدة السؤال قد طابقت كل المعايير. إن مدير البرنامج، الدليل في مشروع تطوير الاختبار، مسؤول عن أمور مثل الحرص على أن خريطة الأسئلة لمستوى المهارات قد تم الإشارة إليها في موضوعات الاختبار، أما المحرر فهو معني أكثر بالتركيز على تحرير صيغة مضمون الاختبار. إن العديد من الموضوعات تم تصديقها أيضاً خلال مرحلة البداية (alpha) أو مرحلة اختبار ما قبل الاختصاص في عملية تطوير الاختبار. أخيراً إن المترجم هو مسؤول عن التأكيد لأن المضمون المُترجم يطابق كل المواصفات المقدمة. تستغرق هذه العملية متوسطياً من 30 - 35 يوماً.

وقد قامت شركة مايكروسوفت بتزويد المترجم بعدة وثائق وأدوات إضافية خلال عملية المركزة ويشمل هذا قائمة لحلول المسائل المطروحة خلال مرحلة ما قبل المركزة. تحوي هذه القائمة أداة مركزة تأكيد النوعية العالمية KIT ودليل الأسلوب وفهارس المنتج. كما ستلاحظون، فإنه على الرغم من أن الأمثلة المقدمة تتعلق بمجال التقنية المعلوماتية، فإن مفهوم خلق وتأمين هذه الأدوات للمترجمين ليس حكراً على مجال IT تتوفر هذه الأدوات بالإنكليزية كما تتوفر باللغة الأصلية.

الجدول 8-2

صيغة تنقيح السؤال (خبير فقط)

الكاتب	PM	محرر	تنقيح	المواصفات
الأسئلة العمومية				
Alpha اختبار ما قبل الاختصاص				
X			X	تم اختبارها من أجل الدقة مقابل البنية الملائمة للمنتج
X	X			خريطة لمستوى المهارة وغاية الأهداف المطابقة
X			X	استعمال مصطلحات تقنية دقيقة تطابق البنية الترقيم المسافات التهجئة والكتابة بحروف كبيرة فمنتج الـ UI والتوثيق.
X		X	X	لا تتضمن ألفاظاً بادئة، أو عناوين في الواجهة، مصطلحات ميزات أو وظائف تم تصميمها أو استعارتها من منتجات أخرى.

الكاتب	PM	محرر	Alpha	المواصفات
X	X		X	تستند على العالم الواقعي، على خبرة العمل.
X			X	لا توجد بشكل حرفي في المواد الموصى بها أو تصديق المنتج.
X		X		معلومات ذات خلفية حالية، حالات حديثة وأهداف في ترتيب زمني.
X		X		تبدأ وتستمر في صيغة الزمن الحاضر.
X		X		استعمل أسماء الشركات والأفراد من قائمة أسماء مصدّق عليها.

الكاتب	PM	محرر	Alpha	المواصفات
X		X	X	تجنب الكلمات غير الموضوعية مثل أفضل، الأفضل وعادةً. استعمل مقاييس موضوعية
X	X	X		هي كاملة وذاتية المضمون بحيث يستطيع الممتحنون الإجابة في شكل مقالة دون رؤية خيارات السؤال.

خيارات السؤال

الكاتب	PM	محرر	Alpha	المواصفات
X			X	تتضمن إجابة صحيحة بحيث تعتبر صحيحة دون شك
X			X	تتضمن عدداً كافاً من الارتباكات، بحيث تكون كلها غير صحيحة ولكنها معقولة بالنسبة للممتحنين الذين يملكون خبرة تقنية غير كافية.
X		X		متناظرة (شبيهة بالتركيب والمضمون) الذي هو على الأقل واحد من خيارات سؤال آخر.

مخططـات

الكاتب	PM	محرر	Alpha	المواصفات
X	X			مرتكزة على قالب VisioNewArtTem-plate.vsd الحالي. (يستبقي الكاتب نسخة مصدر للمخططات للتسليم بعد مرحلة الـ .Alpha
X				محفوظة كـ 16 لون خرائط الأحرف حرف 16 لون حرفي (يحتفظ الكاتب بنسخة خريطة الحرف من المخططات لتسليمها (مرحلة الـ Alpha)
X	X			تم عرضها مع التعابير المطابقة في Word Normal view .

صورة الشاشة

الكاتب	PM	محرر	Alpha	المواصفات
X	X			ليست أكبر من 595 نقطة عالية بواسطة بعد نقطي يبلغ 410
X				محفوظة كـ 16 لون خرائط الأحرف حرف 16 لون حرفي (يحتفظ الكاتب بنسخة خريطة حرف من صور الشاشة لتسليمها بعد Alpha)
X	X			تم عرضها مع التعابير المطابقة في Word Normal view .

الجدول رقم (8-3)

مجموعة مركزة لتأكيد الجودة العالمية IQA مقتطف فقط)

مجموعة مركزة لتأكيد الجودة العالمية الألمانية
المحتويات
1- أسماء البلدان القائمة الأخيرة للأسماء المترجمة للبلدان.
2- مواصفات البلاد مقاييس البلد لألمانية، النمسا وسويسرا.
3- عناوين وأسماء مختلفة هذه القائمة تتضمن أمثلة مختلقة لتجّار Northwint [دخول]، عدة أسماء مختلقة معتمدة وعناوين مثل العناوين الإضافية الألمانية، النمساوية، السويسرية. انظر أيضاً إلى المعلومات المرفقة فيما يتعلق بعناوين IP، التي يمكن أن تكون مفيدة أثناء المركزة أو التصديق.
4- عناصر السطح البيني لصور المستخدم البيانية: تعابير جوهرية إن خريطة الحرف مُشتَملة في ملف رقم المنطقة المثبتة، مخصصة كتنقيح لعناصر وتعابير GUI الأكثر أهمية.
5- مفردات قانونية ومواد مرجعية إن الملفين السابقين تم دمجهم إلى ملف مرجعي واحد لكل الترجمة التي تحتوي على نص قانوني، مثل UELAs، وحق النشر، والنص القانوني في برنامج الحاسوب.

إن أداة مركزة تأكيد النوعية العالمية KIT هي وثيقة شاملة للغاية تم تطويرها في قسم مايكروسوفت العالمي لتطوير المنتجات. مثال عن الموضوعات التي تمت

تغطيتها في Kit IQA مقدم في جدول رقم (8-3) تتألف الـ Kit من جدول من المحتويات مع سلسلة من الوثائق المصدّقة. على سبيل المثال، إذا أراد المترجم أن يعرف ما هي الأسماء والعناوين المختلقة التي يمكن استعمالها في سيناريوهات أسئلة، عندها سيقوم بالنقر على النشرة المطابقة وسوف يرى قائمة أسماء وعناوين كاملة لألمانيا والنمسا وسويسرا؛ لإن مركزة تأكيد النوعية العالمية KIT مصممة لمركزة منتجات برامج مايكروسوفت، فقد لا تغطي دائماً كل الموضوعات التي تنشأ عن تكييف الاختبارات المرتكزة على الحاسوب. لهذا، من الضروري أيضاً أن يتم تزويد المترجمين بدليل أسلوب مشابه للدليل المقدم في جدول رقم 8-4 مثلاً، في هذا الدليل يتم تعريف بعض التعابير مثل buffer، caching، host، names.

إن دليل الطريقة المعروض في الجدول رقم (8-4) مستخدم في حروف العطف مع فهرس المنتج المبين في جدول رقم (8-5) إن فهرس المنتج مفيد بشكل خاص لأنه يشير إلى الصيغة الدقيقة التي يجب أن تستخدم في اللغة الأصلية لكل من التعابير المستخدمة في المنتج. مثلاً إن الترجمة الألمانية "فتح ملف" مدرجة بالقائمة تماماً كما يجب استعمالها.

بعد أن يتم تزويدهم بكل المعلومات، كما هو مبين في الجداول ذات الأرقام 8-3، 8-4، 8-5 يقوم المترجم بترجمة وحدة السؤال النهائية ثم يقوم بعمل نسخة تحريرية على محتوى الاختبار المترجم. بالإضافة إلى وحدة السؤال، فإن المترجم مسؤول عن مقارنة عرض السؤال على الشاشات من المنتج المركز الحقيقي. يتم تقييم خبراء الترجمة استناداً على مقدرتهم بالالتزام بمعايير دقيقة ضمن الإطار الزمني المحدد.

تقوم شركة مايكروسوفت باستعمال أدوات متعددة لتفعيل عملية الترجمة. قد يبدو نموذج لسطح بيني لمستخدم ما على شكل المجموعة التالية من صور الشاشة المعروضة في شكل رقم 8-1 حتى شكل رقم 8-5 إن الشكل رقم 8-1 يمكّن المترجم

من اختيار اللغة واختيار أرقام الاختبار اللازمة في هذا الاختبار، قام المترجم باختيار اختبار حول أساس فن العمارة، وسوف يقوم بترجمته إلى اللغة الفرنسية.

إن الشكل رقم 8-2 يمكّن المترجم من اختيار السؤال ودراسة الحالة التي يجب ترجمتها. في هذا الاختبار، قام المترجم باختيار سؤال متعدد الخيارات 1 (CLEE.1.a) المترافقة مع كل الحالات.

بعد اختيار سؤال ما، يكون المترجم جاهزاً للدخول إلى المحتوى المترجم. يقدم الشكل رقم 8-3 مثالاً عن السطح البيني المستخدم للدخول إلى المحتوى المترجم.

يوضح الشكل رقم 8-4 الآلية المستخدمة من قبل المترجمين للقيام بتغييرات عالمية للاختبار عبر وحدة السؤال بأكمله.

يوضح الشكل رقم 8-5 القدرة على مشاهدة كيف سيبدو السؤال بالضبط بالنسبة للممتحن في كلا اللغتين الإنكليزية واللغة المترجمة. من أجل التمكن من رؤية الاثنين معاً، يمكن للمترجم إما أن يقوم بترتيب الشاشات بتحريكها أو باستعمال الأمر "ترتيب الكل" تحت قائمة الخيارات، WINDOW.

الجدول 8-4

دليل أسلوب المصطلحات التقنية المعتمدة على برنامج (Windows NT 4.0 مقتطف)

Account Policy	خطة الأمان لـ Windows NT التي تتحكم بكيفية استعمال كلمة السر بواسطة حسابات المستخدم
Broadcast message	رسالة عبر الشبكة مرسلة من كمبيوتر ما وموزعة لكل الأجهزة الأخرى على نفس الجزء من الشبكة كالحاسوب المرسل.
Buffer	حجم احتياطي من الذاكرة حيث تحفظ البيانات بشكل مؤقت قبل طباعتها
caching	طريقة تستخدم من قبل مخدمي الاسم DNS لتحسين الأداء. كما تستلزم عملية مخدمي الاسم DNS، فهم يحفظون المعلومات بشكل مؤقت في مخزن محلي (cache) ويستعملونها للإجابة عن الاستفسارات الإضافية للمعلومات نفسها.

Catalogs	معلومات حول كل مجموعة من نظام حفظ البيانات وموقعهم المخزن على مجموعة من الشرائط. تتضمن معلومات الـ catalog عدد الأشرطة في مجموعة من الشرائط كما تتضمن التاريخ الذي تم إنشاء هذه الشرائط وتواريخ كل ملف في هذا الـ Cata-log وهي تخزن على آخر شريط في المجموعة. وهي تنشأ لكل مجموعة من نظام حفظ البيانات.
Discovery	محاولة من قبل خدمة Net Logon المتعلقة بالحاسوب لتحديد حقل
Domain controllers	الحواسب التي تدير مخدم Windows NT الذي يشارك معلومات أساسية دليل لمخزن الأمان ومعلومات حساب المستخدم لكل الحقل.
Full synchronization	يرسل الـ PDC نسخة إلى المعلومات الأساسية لدليل بأكمله إلى PDC
Host name	جزء من البنية الاسمية DNS وهو يشير إلى أداة محددة تربط بشبكة الانترنت TCP/IP في اسم ملف صالح بشكل كامل (FQDN)، هو الجزء الأكبر الباقي (إن مجموعة الحروف قبل المرحلة الأولى) من الاسم.
Host	الحاسوب الأساسي في نظام الحواسب أو المحطات الموصولة بروابط اتصالات.

الجدول 8-5

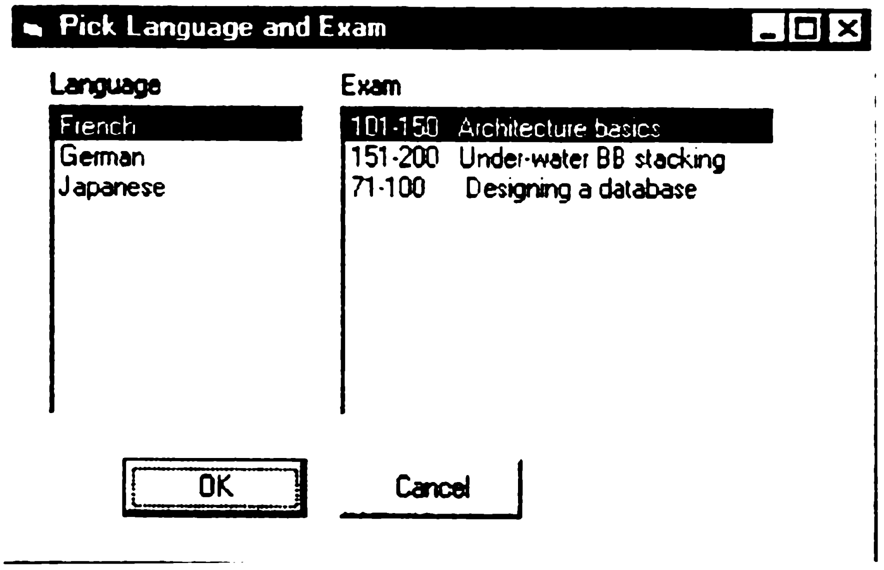
مفردات المنتج (مقتطف فقط)

English	*German*	*Type*	*Name*	*Product*
# File Opens	Dateien geöffnet	COM	NTLANUI.DL_	NT
# Opens%0	Öffnungen%0	TXT	NETMSG.DL_	NT
#10: S=Specify Additional SCSI Adapter (10029)	#10: Z=Specify Additional SCSI Adapter (10029)	TXT	USETUP.EX_	NT
#12: O=Overwrite (10065)	#12: U=Overwrite (10065)	TXT	USETUP.EX_	NT
#16: U=Continue Upgrade (10087)	#16: A=Continue Upgrade (10087)TXT	USETUP.EX_	NT	
#17: Y=Yes, I agree (10089)	#17: J=Yes, I agree (10089)	TXT	USETUP.EX_	NT
#Programs#*.exe;*.pif;*.com;*.bat;*.cmd#All files (*.*)#*.*#	#Programme#*.exe;*.pif;*.com;*.bat;*.cmd#Alle Dateien (*.*)#*.*#	TXT	SHELL32.DL_	NT
$* Symbol replaced by everything following macro name on command line.	$* Symbol für alles, was auf der Befehlszeile nach dem Makronamen folgt	TXT	AUTOCHK.EX _AUTOCONV. EX_,ULIB.DL_	NT

ما بعد المركْزَة Postlocalization

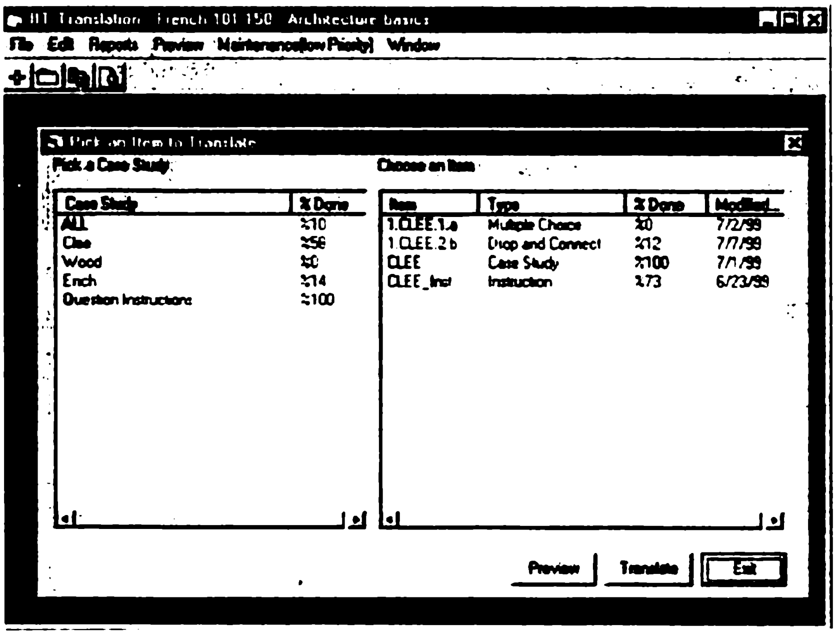
بعد أن يتم تحضير أسئلة الاختبار من أجل الاستعمال في لغة ثانية، يتم إنشاء قرص لتنقيح الاختبار. يقوم مدير المركزة المشتركة بإرسال القرص إلى ممثل شركة مايكروسوفت في البلد الأصلي وذلك بغرض تنقيح تقني من قبل متكلم أصلي للغة الذي يكون أيضاً بليغ باللغة الإنكليزية. إن الهدف من التنقيح التقني هو للتأكيد على أن محتوى الاختبار مطابق للمعايير المحددة المتعلقة بقياس وسرعة ودقة

العمليات العقلية (كما نوقش سابقاً في هذا المقطع)، الدليل التقني والثقافي المحدد من قبل شركة مايكروسوفت. لا يقصد بهذا أن يكون تنقيحاً لغوياً، على الرغم من أنه في حالات معينة، تكون المساهمات اللغوية ضرورية. هناك مثال مبين عن السطح البيني المستخدم لإكمال هذه المهمة في الشكل رقم 8-6 في هذا المثال لسؤال باللغة الألمانية يتم التزويد بصندوق حوار/ Window وذلك للقيام بالتعليقات (Kommentare) على السؤال.



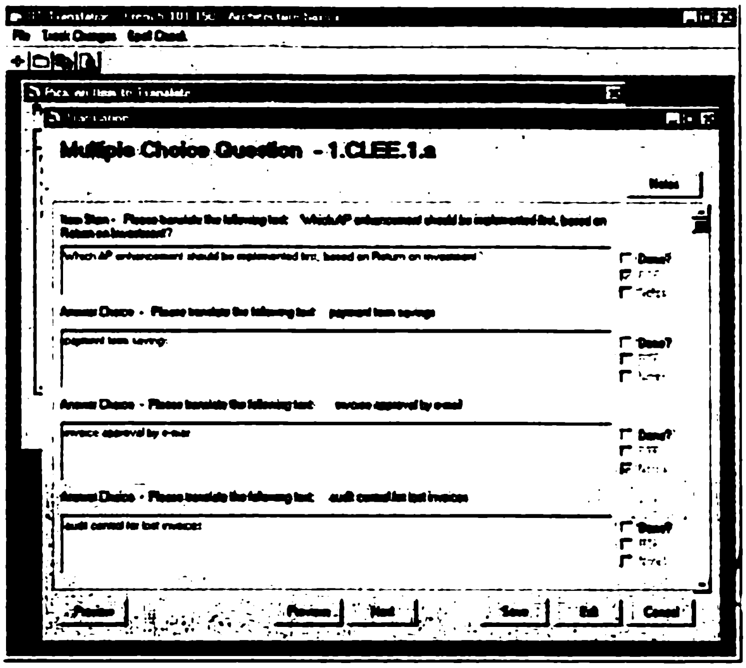
الشكل رقم 8-1

صور الشاشة لاختيار اللغة



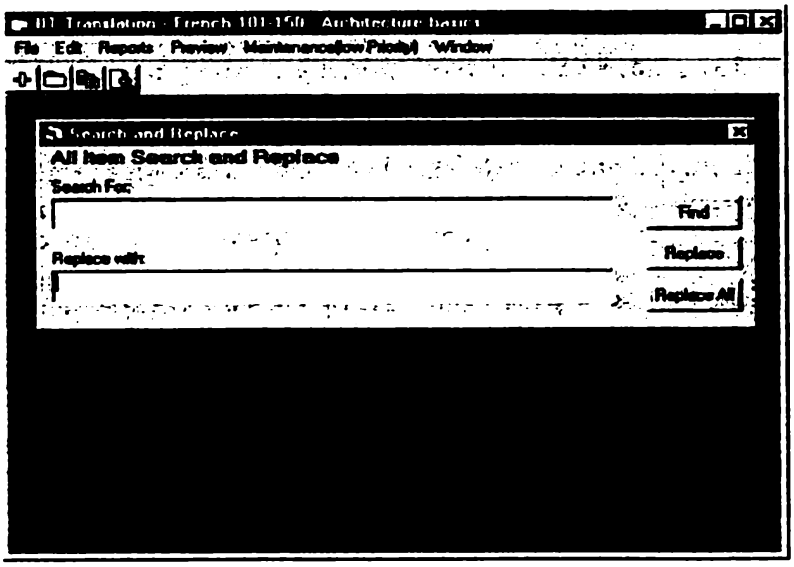
الشكل رقم 8-2

صور الشاشة لاختيار التعابير



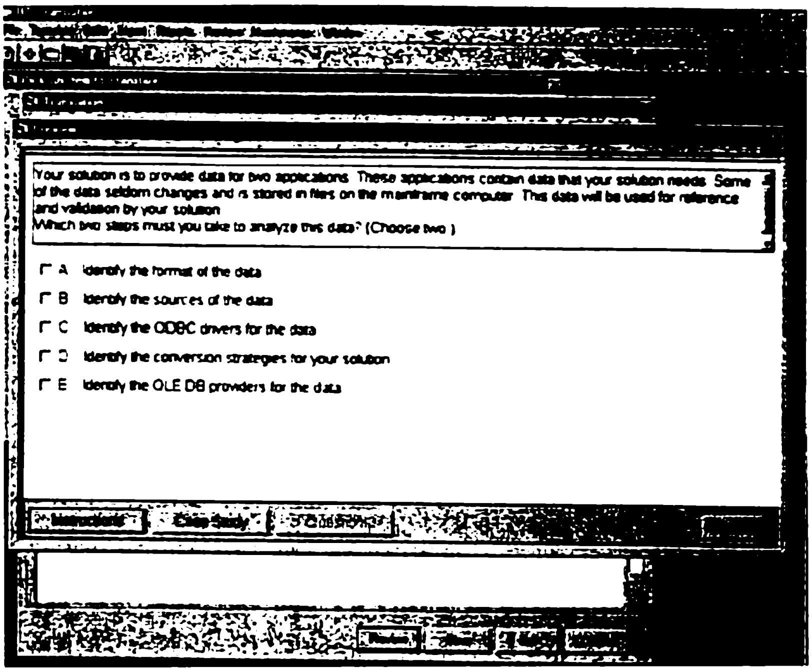
الشكل رقم 8-3

صور الشاشة لترجمة التعابير



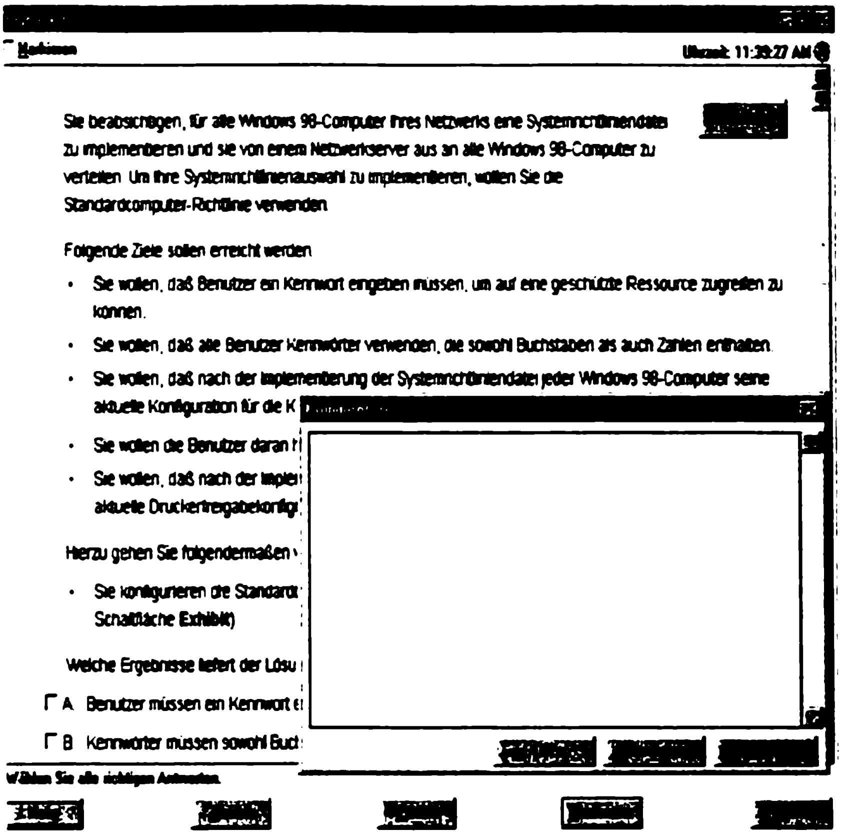
الشكل رقم 8-4

صور الشاشة للبحث والاستبدال



الشكل رقم 8-5

صور الشاشة لتنقيح التعبير



الشكل رقم 8-6

صور الشاشة لتعليقات المنقح

إن التنقيح التقني هو الآن أكثر أهمية بالنسبة لصدق الاختبارات المصدّقة التابعة لشركة مايكروسوفت، بحيث إن هذه الاختبارات تتضمن أسئلة تصورية تتعلق بالمنتج. حين يتم إرجاع تعليقات المنقح التقنية، يقوم المترجم بإعادة النظر بالتغيرات والمساهمات اللازمة على التغيرات التي تم الجدال حولها من مساعد مدير المركزة، كما تقتضي الضرورة. تُدرج كل التغيرات الملائمة الموصى بها من قبل المنقح التقني إلى الصيغة النهائية للاختبار، حالما يتم الموافقة عليها من قبل كل الأطراف. ثم يتم بعد ذلك نشر الاختبار ويتم جعله متوفراً في كل أنحاء العالم بلغات متعددة. حين يتم تسليم أمثلة كافية عن أسئلة الاختبار المركز، يعمم تقرير متعلق بالإدارة الداخلية بشأن الاختبار.

في حـال وجـود أية أسـئلة لا تفي بـالغـرض فـإنـه يتم إنجـاز كل من التـحليل و التحليل الكمي (لمثال عن نوع التحليلات التجريبية التي قامت شركة مايكروسوفت بتنفيـذها، انظر إلى روبن، سيرسي وهـامبلتون، 2003) تم جمع هذه المعلومات من خلال تعليقات الاختبار ضمن الاختبار. بالإضافة إلى تصاعدات الاختبارات المرسلة إلى شركة مايكروسوفت في شكل رسائل إلكترونية وفاكسات ورسائل. علاوة على ذلك تقوم شركة مايكروسوفت بإجراء دراسات وظيفية تتعلق بأسئلة تفاضلية (مونيز، هامبلتون وإكسنغ، 2001، روبن وآل، 2003، سيرسي، فيتزجيرالد وإكسنغ، 1998) على اختباراتهم الأكثر شعبية وذلك لتحديد إذا كان هناك لزوم لتحديد اختبار ما.

كما هو الحال في أي شيء بشركة مايكروسوفت، فإن الأشياء دائمة التغيّر بخطوات سريعة. تاريخياً اعتمدنا على مترجم واحد فقط. أما الآن فقد يتم استخدام واحد أو اثنين، كما توصي هيئة دلائل تكييف الاختبار ITC باعتبار أنه تم وضع تأكيدات أكثر على الحقيقة التجريبية لعملية المركزة، فقد بات السؤال الآن هل توجد الفروقات عبر اللغات بسبب اختلافات مضمون محدد أو بسبب اختلافات أكثر عمومية؟ أو هل توجد الاختلافات بسبب مشكلات في الترجمة أو التمرين على الاختبار؟ إنه لمن المهم أن نتابع البحث عن اتجاهات وأن نقوم بتغذية استرجاعية للنتائج وذلك لتحسين العملية بأكملها. لقد اكتشفنا أن إضافة أدوات أخرى وحواجز اختبار لكل العملية تعمل على الإضافة إلى النجاح في الترجمة النهائية.

شكر

يوجه تقدير خاص وامتنان إلى كل من آن ميري مكسويني، وآنجيلا جونسون، وشركة مايكروسوفت لمساعدتهم في إعداد العمليات التي تم شرحها ضمن هذا الفصل.

*****

المراجع

translation or the training? It's important to continue to look for trends and to feed the results back into improving the entire process. We've found that adding additional tools and checkpoints up front in the process adds to the success in the final translation.

ACKNOWLEDGMENTS

Recognition and appreciation is given to Anne Mrie McSweeney and Angela Johnson, Microsoft Corporation, for their assistance in developing the processes described within this chapter.

REFERENCES

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K., Sireci, S., & Robin, F. (1999). Adapting credentialing exams for use in multiple languages. *CLEAR Exam Review, 10*(2), 24–28.

McKillip, J., & Cox, C. (1998). Strengthening the criterion-related validity of professional certifications. *Evaluation and Program Planning, 21*(2), 191–197.

Mills, C. N., Potenza, M., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1,* 115–135.

Robin, F., Sireci, S., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing, 3*(1), 1–20.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovatins in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362.

9

تحويل مقياس فكسلر لقياس ذكاء البالغين: محاولة تكييف اختبار مبكرة ذات نتائج قيِّمة

كارلوس ي. مالدونادو
بوتنام/وستشستر

كيرت ف. كيسنجر
جامعة سانت توماس

في بعض المراحل ترجمت/ كيفت بعض الاختبارات النفسية من لغة إلى أخرى كي تستطيع الاستخدامات الدولية القيام ببعض الدراسات مثل المقارنات الثقافية على سبيل المثال (انظر غريس، 2003، هامبلتون، 2002). في بعض الحالات تحدث تلك التكييفات لأن صدق البنية لإحدى الأدوات أو إحدى الصفات النفسية معروفة في لغة و وثقافة واحدة فقط ويأمل البحث في استخدامها في لغة أو ثقافة أخرى باستخدام إما الأداة ذاتها أو على الأقل إحدى الأدوات المشابهة للأداة الأصلية بما أن بعض الدول مثل الولايات المتحدة أصبحت متعددة القوميات لذلك هناك حاجة لأدوات نفسية ذات قيمة مثبتة في لغات غير الإنكليزية إلى حد كبير حتى داخل البلاد كيسنجر، (1992). غالباً عندما لا تتوفر نسخ اختبار محترمة أو تقاويم في اللغة الإسبانية يستخدم علماء النفس الذين يحتاجون إقامة قياسات نفسية ما يمكن اعتباره أسلوباً غير رسمي لإقامة التقويم النفسي:

بعض الظروف التي لاحظناها عند استخدام مقياس فكسلر لقياس ذكاء البالغين (WAIS) هي:

أ - استخدام الأداة في اللغة الإنكليزية ومحاولة أخذ الاختلافات في اللغة بالاعتبار عند تفسير الدرجات.

ب- استخدام أداء جزء من الاختبارات (d) واستخدام إرشادات في اللغة الإنكليزية أو الإسبانية.

ج- استخدام مترجم فوري، أو (d) إحالة إدارة الاختبار إلى زميل ناطق باللغة الإسبانية أو إلى مساعد يستطيع ترجمة الإرشادات وبنود الاختبار في أثناء إقامته.

إن التقيد بتلك الإجراءات غير مرض وفي بعض الحالات غير أخلاقي (لوبيز وروميرو، 1988ص. 264).

بالرغم من أن المشكلات التي تتضمن مثل تلك الاستخدامات غير الرسمية للاختبارات قد تحدث في لغات عديدة، لكن من المحتمل حدوثها في اللغة الإسبانية في الولايات المتحدة، خاصة فيما يدعى "مخاطرة عالية" في ظروف الاختبار مع ملاحظة الحجم الكبير والمتنامي لتلك المجموعات السكانية.

"إن إقامة الاختبارات النفسية في اللغة الإسبانية أكثر ملاءمة من أي لغة أخرى ما عدا اللغة الإنكليزية في الولايات المتحدة؛ إن معطيات الإحصاء السكاني عام (1990) يشير إلى أن 10٪ من عدد السكان الكامل هم إسبان مقيمون في الولايات لمتحدة ويقيمون بشكل مركز في فلوريدا، نيويورك، كاليفورنيا، وتكساس". (دمسكي، ميتتبيرغ، كوبتار، كاتل و غولدن، 1998، ص115). كانت معظم جهود تطوير الاختبارات لتكييف الأداة في اللغة الإنكليزية إلى اللغة الإسبانية تهدف إلى معايير لاستخدامها مع الأطفال (لوبيز وروميرو، 1988، ماكشين وكوك، 1985) في

كتابات ماكشين وكوك حول استخدام الترجمة الإسبانية لأداة فكسلر، تم إقامة سبعين دراسة على الأكثر، منها اثنتان فقط ناقشت تقويم البالغين.

إن (EIWA) هي التكييف إلى اللغة الإسبانية لمقياس فكسلر لقياس ذكاء البالغين (WAIS) تم نشره كغيره في أدوات الذكاء لفكسلر، من قبل جمعية علماء النفس. صدر مقياس (EIWA) عام 1968 بعد تكييفه من نسخة اللغة الإنكليزية إلى الإسبانية (مع التغيرات المرافقة للتأكد من التكافؤ الثقافي مع إعادة المعايرة بشكل كامل في بورتوريكو. جرى تعديل بعض البنود في أجزاء الاختبار الثانوية، أو حذفت أو زيدت لجعل الاختبار موثوقاً أكثر للقياس السيكولوجي، ومناسباً وملائماً ثقافياً لشعب بورتوريكو وشعوب لاتينية أخرى. في أكثر الأقسام احتفظ (EIWA) بالبنية الأساسية في أجزاء الاختبار الفرعية (غرين 1964). بعض حالات تكييف (WAIS)لتكوين (EIWA) جرت الإشارة إليها لاحقاً في هذا الفصل، انظر غرين ومارتينز (1967) للحصول على وصف كامل. أحرز (EIWA) مستوى استخدام واسع في الممارسة التطبيقية، لكن الأبحاث عن فعاليته كانت محدودة حتى وقت قصير. كشف بحث استخدم معطيات سيكولوجية في برامج كمبيوتر من عام 1967 حتى تموز 1993 عن 12 دراسة، طبعت بعضها وبعضها الآخر لم تطبع، تعالج النسخة الإسبانية لاختبار (WAIS) إحدى تلك الدراسات (كوند لوبيز ودومينكا لوبيز، 1977) لا تتعلق مباشرة بالدراسة التي نجريها لأنها تنقد الترجمات، التي تستخدم حالياً في إسبانية، وشسلر الأصلية و (WAIS) لكنها لا تذكر EIWA منذ تقديم EIWA عام 1968 جرت 11 دراسة فقط عليه.

عندما تم استخدام EIWA لأول مرة كان المقياس الوحيد للقياس السيكولوجي الذي بدا ملائماً والمعيار المطبوع الذي يمكن استخدامه للتقويم الذهني للمجموعات من أصل لاتيني ويبقى واحداً من المعايير القليلة لهذا الغرض حتى الآن. بسبب مركزه الذي لا يمكن منافسته أصبح EIWA الأداة الرئيسة للتقويم الذهني لتلك

المجموعات في الولايات المتحدة بما فيها بورتوريكو. كان ولا يزال يستخدم بشكل واسع في اتخاذ القياس السيكولوجي في المواقف الحرجة.

لأن السكان المنحدرين من أصول لاتينية هم الأقلية الأكثر سرعة في النمو في الولايات المتحدة آيد، 1992، ماكياس، (1977)؛ لذلك من الطبيعي أن يتوقع المرء أن يثير EIWA اهتماماً أكثر بالاختبار وبالأبحاث التي تجرى لإثبات صدق وثبات ذلك الاختبار. بالرغم من الاستخدام الدائم لـ EIWA فإن الباحثين تجاهلوا حتى وقت قريب مقارنة بأبحاثهم عن مقياس فكسلر في اللغة الإنكليزية.

الفروق في الدرجات عبر اللغات:

في بداية استخدام EIWA، لاحظ علماء النفس الناطقون باللغة الإسبانية بشكل غير رسمي أنه يعطي غالباً محصلات ذكاء عالية للأفراد عند مقارنتها بدرجات معايير في اللغة الإنكليزية أو عندما تقدر الإمكانيات المعرفية المطورة في مستويات وظيفية معروفة. إن الطريقة الرسمية التي نقلت فيها تلك المشاهدات مسؤولة إلى حد ما عن عدم وجود معلومات كافية ومصادر ممكنة لتلك الفروق. حتى وقت قصير، وبعد 20 عاماً من نشر EIWA، بدأت المقالات في المجلات الدورية بتحسين ذلك الموقف وذلك بنشر معلومات ذات علاقة باستخدام (EIWA لوبيز وروميرو، 1988). توجد في الوقت الحاضر كثير من الآراء التي تفسر الاختلافات المزعومة في الدرجات. من الناحية الأولى من المعقول أن تعكس الدرجات العالية لـ EIWA الحساسية الثقافية للاختبار تجاه الأشخاص ذوي الأصول اللاتينية وبذلك لا تمثل تضخم الدرجات ولكنها تمثل تقديرات جيدة للذكاء أكثر من تلك التقديرات في نتائج الاختبار. إن هذا الظن تسببه الفكرة القائلة إن EIWA اختبار متحيز وغير عادل عند استخدامه لأشخاص لغتهم الأم ليست اللغة الإنكليزية. من ذلك المنطلق فإن الدرجات العالية تمثل تقديرات صادقة عن مستوى الطلاب، الذين يجرون الاختبار الوظيفي، وبذلك يكون الرأي التطبيقي من ذلك

المنظور حول المستوى الوظيفي ومستوى الذكاء ذا صدق للاستخدام لمجموعة ولغة محدودة تمت مراقبتها. على كل حال، أعطت دراسة أولية لصدق بنية EIWA الانطباع أن ذلك لا يمكن أن يكون صحيحاً غوميز، بيدمونت وفليمنج، (1992).

من جهة أخرى يستطيع المرء أن يشرح الاختلافات بين حاصل الذكاء المقدر لمستوى وظيفي معروف وحاصل ذكاء تم الحصول عليه من الاعتماد على قياس سمات وصفات نفسية مختلفة إلى حد ما. إن مؤيدي ذلك الرأي يعتقدون بأن ترجمة البنود جميعها تقريباً واستبدال بعضها يؤدي إلى تغيير الاختبار عن مضمونه الأصلي ولا يعيد تقديم مواصفات قياس WAIS، بل يؤدي إلى تداخلها.

هناك إمكانية ثالثة لشرح التفاوت الواضح في الدرجات وهو أن رفض ادعاء صدق EIWA في بعض الأجزاء أو الأماكن هو السبب في اختلافات الاختبار نفسه. إن مؤيدي ذلك الرأي هم الذين لاحظوا أن EIWA قد جرى إعطاؤه بشكل سحري الصدق المفترض؛ لأنه قد جرى تطويره من WAIS ذي السمعة العالية. من ذلك المنظور لا تعود الاختلافات في الدرجات إلى اختلافات في المجموعات ولكن إلى قواعد خاطئة، معايرة الاختبار، مقاييس الاختبار الفرعية، بنود الاختبار والمواد، عملية بناء الاختبار/ تكييفه/ وتحليل البنود. إن الرأي المعدل لتلك الإمكانية الثالثة هي أن الاختبار صادق ولكن يجب موازنته بطريقة معقولة مع WIAS الذي قام عليه. بعض تلك الإمكانيات لحسن الحظ يمكن أن يجري تأكيدها بشكل تجريبي.

موضوعات عن نموذج معايير EIWA:

ما هو الدليل على وجود مشكلات متأصلة في EIWA؟ لسوء الحظ قليل من الدراسات المطبوعة تعالج ذلك الموضوع مباشرة. على كل حال، فإن الفحص الدقيق لكتيب EIWA والتقارير الفنية تكشف تفاصيل مفاجئة. على سبيل المثال: أحد العوامل السهلة للتأكيد هو عما إذا كان هناك تناقض في معايرة النموذج بالمقارنة مع الإحصاء السكاني في بورتوريكو عام 1960. مثل WAIS استخدم EIWA

إجراءات كأخذ عينات مقسمة إلى طبقات لصفات مجموعات ستة: العمر، الجنس، الإقليم الجغرافي، المهنة، مكان الإقامة (المدينة، ريف). كان الفرق (أبيض-ليس أبيض) مراقباً في WAIS ولكنه لم يراقب في EIWA إن النموذج المعاير استخدم في بورتوريكو فقط.

قام بعملية تحليل المقارنة بين معطيات الإحصاء السكاني 1960، ومجموعات EIWA المعايرة في العوامل الستة لوبيز وروميرو (1988) كجزء من عملية البحث الحالي. بالرغم من أن كتيب EIWA يبين توزع تلك العوامل ونسبتها في اثنين أو أكثر فقط كل مرة، كان ممكناً، بعد أخذ حجم النموذج المعاير (604 نساء و 523 رجلاً) إقامة جدول حسب التواتر وعزل العوامل التطبيقية الستة.

يبين التحليل أن نموذج عامل الجنس كان من عدد السكان، ولكن بالرغم من أن كتيب EIWA يقرر أن (معيار النموذج يمثل سكان بورتوريكو بشكل كاف) ص8 فإن هذا التحليل يظهر أن العمر، الإقليم الجغرافي، العمل والثقافة أظهرت اختلافات ملحوظة ($P<0.001$)* في القيمة الأصلية السكانية. إن أهمية الاختلافات في عوامل السن ربما أقل في عوامل المنطقة السكانية. المهنة، والثقافة: لأن أداء الطالب في EIWA يقارن مع قواعد العمر فإن أهمية الحصول على نموذج معاير يحدد فيه العمر في نسبة القيمة الأصلية للسكان ليست ضرورية للنتائج. إن العدد القليل أو الكثير في مجموعة منظمة حسب العمر تؤثر على مصداقية النموذج ضمن إطار عمر محدد.

إن الاختلافات في النماذج التي أخذت حسب المناطق السكانية في بورتوريكو، بينت أن المهنة والثقافة كانت لهما أهمية إحصائية ويجب النظر إليهما على أنهما عاملان ذوا أهمية محتملة. عموماً تبين المعطيات أن المدن الموجودة في مركز الجزيرة وحتى شمال الساحل (الإقليم الثالث) لم يجر تمثيله كما يجب. هنالك

(*) مصطلح يعني أن احتمال حدوث النتيجة يكون أقل من مرة في كل 1000 مرة. ويسمى «احتمال الخطأ».

فروق مهمة في تلك المناطق في بورتوريكو فيما يتعلق بالثقافة ومتغيرات الوضع المادي/ الاجتماعي. كما أن التوزع المهني أظهر أن الأكثرية هم من العمال، ربات المنازل، الطلاب العاطلين عن العمل، بينما أظهر أعداداً أقل من الذين صنفوا (الآخرين) في الإحصاء السكاني المهني. إن تأثير تلك الاختلافات من الصعب تقويمها. على كلٍ إن الذي يمكن أن يكون له معنى أكبر هو الاختلافات الثقافية الناتجة عن مشاركة عدد قليل من الطلاب الذين أنهوا دراسة ثماني سنوات وعدد أكبر أنهوا دراسة تسع سنوات أو أكثر في الاختبار؛ وكنتيجة لذلك كان معيار النموذج أكثر من المتوسط في مستوى الثقافة. ذكر واحد أو اثنان في تلك الموضوعات في الكتيب بشكل مختصر وفي التقارير اللاحقة ولكن الاستنتاج النهائي للذين أجروا عملية التكييف كان أن النموذج مناسب إلى حد ما. وفي الحقيقة فإن كتيب (EIWA فكسلر، 1968) يذكر فقط التباين الذي وجد في المستوى الثقافي، ويمضي بالقول إن ذلك الخطأ يتوافق مع النزعة السكانية المعروفة، أي أن مستوى الثقافة في بورتوريكو بدأ بالازدياد في الستينيات بعد جمع معطيات الإحصاء السكاني.

استخدمت دراسة أطروحة لأحد أعضاء البحث في EIWA مجموعة معطيات المعيار هيرانس، (1969) ليبحث العلاقة بين درجات EIWA وعدة متغيرات منها الجنس، الثقافة والمناطق السكنية في بورتوريكو. وجدت أن أداء الذكور كان أفضل من أداء الإناث في الأداء اللفظي، درجات المقياس التامة بصرف النظر عن الثقافة أو المنطقة السكنية. وجدت هيرانس أيضاً، كما توقعت، أن الممتحنين في مناطق المدن سجلوا درجات أعلى من طلاب المناطق الريفية بصرف النظر عن الجنس أو الثقافة. أخيراً وجدت أنه كلما كان الممتحن متعلماً أكثر يكون حصوله على درجات أعلى أكثر دون النظر إلى جنسه أو منطقة سكنه. تتوافق نتائج تلك الدراسة وتؤكد أهمية الاستنتاجات التي ذكرت سابقاً فيما يعلق بعدم ملاءمة المعايير المحتملة بسبب نماذج منقوصة. راجع لوبيز وروميرو (1988) تلك المعطيات واعتبرا أن أمور المعايرة لها أهمية ذات مغزى في تفسير درجات EIWA.

موضوعات عن صدق التركيب/ البنية:

قام غوميز وزملاؤه (1992) بدراسة تحليل عوامل EIWA أخضعت دراستهم كلاً من EIWA و WAIS على تحليل عناصر رئيسة لكي تستطيع اختبار إمكانية مقارنتهم (استخدموا درجات التطابق). أشارت نتائجهم أنه بالرغم من اختلاف المحتوى في كلٍ من الاختبارين، "فإنه EIWA، في التركيب على الأقل، يظهر وكأنه انعكاس قياس سيكولوجي لـ WAIS" (ص. 320). تؤكد لنا تلك الدراسات أن الأبعاد الذهنية الأساسية التي جذبت انتباه EIWA هي شبيهة بتلك التي جذبت انتباه WAIS. وعلى كلٍ لم تتناول تلك الدراسة موضوعات المقارنة بين معايير كل منهما.

استخدم مارتينيز- يوروتيا وسبيلبيرغر (1973) EIWA. في محاولتهم لقياس العلاقة بين حالة وسمة القلق وبين الذكاء. أشرفوا على إدارة اختبار EIWA والنسخة الإسبانية من اختبار حالة/ سمة القلق لسبيربرغر (STAI) على 40 مريضاً نفسياً في مستشفى سان خوان للمحاربين القدماء. وقد تأكدت توقعاتهم على أن قياسات حالة وسمة القلق تتعلق سلبياً بالأداء في EIWA. لسوء الحظ لم يذكر الباحثون توزع درجات EIWA في دراستهم. على كلٍ صرّح أحدهم (مارتينيز- يوروتيا، في حديث خاص، 26 كانون الأول، 1984 أنه بالرغم من أن مجموعة المتحنين كانوا مرضى نفسيين فإن متوسط معدل الذكاء في EIWA كان قريباً من 120، ولم يستطع مارتينيز تفسير ذلك المعدل العالي نسبياً. بالرغم من أن ذلك لا يعد استنتاجاً ولكن تلك الملاحظة تتوافق مع تجربة الأطباء الذين عادة يحصلون على درجات في EIWA أعلى من المتوقع.

موضوعات متعلقة بترجمة وتكييف الإدارة

وصفت هيرانس (1973) بعض الترجمات والموضوعات الثقافية المتعلقة بتكييف EIWA وأفادت أن المسؤولين عن عملية التكييف استخدموا أكثر اللغة الإسبانية شموليةً وتجنبوا استخدام اللغة الإقليمية لكي يمكن استخدام EIWA

أكثر سهولة في البلاد الناطقة باللغة الإسبانية بالإضافة إلى بورتوريكو. وللتأكد من ذلك تم إرسال كتيب الاختبار المترجم إلى عدة لغويين في ثلاثة أو أربعة بلدان في أميركا الجنوبية لملاحظاتهم فيما يختص بالصياغة. على سبيل المثال في بعض أقسام الكتيب يطلب من الممتحنين استخدام كلمات متماثلة تستخدم في بلدهم فقط - وهذا إجراء غير طبيعي عند إقامة اختبار قياسي. أفادت هيرانس أيضاً أن كثيراً من البنود قد تم تغييرها أو حذفها لأنها، حسب رأي أعضاء الهيئة في بورتوريكو، لم تكن صالحة لشعب بورتوريكو لأنها لا تمس العالم الواقعي في بورتوريكو. "على سبيل المثال البند # 18 وهو إكمال جدول في جزء من اختبار WAIS تم حذفه لأن الثلج لا يسقط في بورتوريكو"(ص28). ولكنها اعترفت أن إجراء الاختبار الأولي انحصر في مطوري الاختبار، بعض الطلاب من جامعة بورتوريكو، ومجموعتين صغيرتين. وضعت مستويات الصعوبة في البنود حسب تلك المجموعتين الصغيرتين.

ركزت دراسة ثانية لميليندز، (1994) على الاعتبارات الأخلاقية لـ EIWA. وثّق ميليندر عدة موضوعات جدية عن الإدارة/ EIWA التي تجعل مقارنة الدرجات مع WAIS صعبة في أحسن الأحوال أو شبه مستحيلة كما أفاد عند مقارنته WAIS و EIWA.

"إذا كانت أوجه الإحصاءات والعوامل متشابهة في هذين الاختبارين، فإن المحتوى، والدرجات ونتائج أخذ واحد من الاختبارين مختلفة بشكل كبير. على المرء أن يتوقع أن ترجمة جيدة عبر الثقافات/ اللغات ينتج عنها بعض التغييرات في بنود الاختبار خاصة في تلك التي بها خاصية ثقافية. ولكن التغييرات التي وجدت في EIWA كانت عامة حتى إنها تخطت أي تعديل ثقافي معقول وذلك ليس بتغير محتوى الاختبار فقط بل بتغير المدة الاختبارية، وإلغاء بعض الدرجات. كل تلك التغييرات جعلت EIWA اختباراً أكثر تساهلاً حتى إن بعض الأجوبة التي سجلت كخطأ في اللغة الإنكليزية سجلت صحيحة في اللغة الإسبانية. ويجب أن يكون هناك أي سبب ثقافي أو غير ثقافي لإعطاء درجات لإجابات خاطئة على أنها صحيحة (ص. 389)".

لاحظ ميليندز (1994) أنه في بعض الحالات فإن طبيعة الأسئلة في WAIS تتطلب بعض التغييرات في EIWA لأسباب ثقافية بحتة. هذا يتضمن تقريباً كل المفردات، المعلومات، وأسئلة الفهم. على سبيل المثال: أحد أسئلة WAIS يطلب من الممتحنين تسمية أربعة رؤساء في الولايا ت المتحدة منذ عام 1990 تم استبداله بسؤال عن ثلاث لغات تستخدم في الولايا ت المتحدة (تعطى العلامة التامة إذا سميّت اثنتين فقط). عموماً في قسم اختبار المعلومات في WIAS هناك 27 بنداً ويتوقف الفاحص عن الاختبار بعد خمسة أغلاط متكررة بينما يحتوي اختبار معلومات EIWA 32 بنداً ويستمر الفاحص في الاختبار حتى حدوث سبع أخطاء متكررة. ووثّق ميليندز أن الإجابات المترجمة لبعض الأسئلة تعطى درجة واحدة في EIWA بينما لا تعطى أي درجة في WAIS. على سبيل المثال: إذا استطاع طالب في اختبار اللغة الإنكليزية تسمية اثنين من ثلاثة من الأوردة الدموية المطلوبة لا يعطى أي درجة بينما الطالب في اختبار اللغة الإسبانية يعطى لهذا الجواب نقطة واحدة.

وثّق ميليندز أيضاً عدة موضوعات عن الدرجات التي تربك المقارنة بين المقياسين. "إذا استطاع أحد أن يردد ستة أرقام إلى الأمام وخمسة إلى الخلف في الإنكليزية فإنه سيحصل على درجة مقياس 10" (ص. 390). إذا تم أداء ذلك في اللغة الإسبانية في EIWA فإن مقياس الدرجة هو 14 من الواضح أن السلوك نفسه يجب أن ينتج ذات مقياس الدرجة من منظور مفهوم الدرجة. ناقش ميليندز أن تضخم الدرجات موجود في كل الاختبارات الفرعية وسبب المشكلة أن معيار الاختبار هو متوسط 100 في بورتوريكو لأعلى موازنة معدل WAIS وقد اعتقد أن EIWA يغالي في تقديرات حاصل الذكاء مقارنة بـ WAIS حوالي 20 نقطة في المستوى الأدنى والمتوسط لتوزيع الدرجات و12 نقطة في المستويات العالية؛ وذلك يعود إلى عامل الحد الأعلى للدرجات على الأغلب.

دراسات صدق متلازمة/ مقارنات
مع النسخ الإنكليزية لاختبارات فكسلر

ما يدعو للدهشة أن عدداً قليلاً من الأبحاث حتى هذا التاريخ حاولت أن تتقصى الصدق المتلازمة لـ AIWA وذلك بمقارنته بأدوات أخرى لفكسلر. تقصى ديفيز وورودريغوز (1979) صدق EIWA مع معطيات لا تعتمد على تلك المعطيات التي تم جمعها من المجموعات المعيارية الأصلية. تم إقامة تلك الدراسة في وحدة المعالجة الدائمة في مركز الصحة العقلية في منطقة القناة. استخدمت خطتا بحث. في الخطة الأولى جرت مطابقة نماذج (N = 14 each) من الممتحنين الذين يتكلمون اللغة الإنكليزية مع آخرين يتكلمون اللغة الإسبانية حسب العمر، الجنس والثقافة في مرحلة القبول الأولية وقد قاموا بإجراء واحد من الاختبارين EIWA أوWAIS للمفردات وللاختبارات النفسية العلمية الفرعية. جرى الإشراف على تلك الاختبارات من قبل فاحصين ثنائيي اللغة لمجموعة EIWA وفاحصين ناطقين باللغة الإنكليزية لمجموعة WAIS مستخدمين الإرشادات القياسية. قدرت المقاييس التامة، الأداء والدرجات اللفظية من درجات الاختبار الفرعية المتوفرة. أظهرت النتائج أن المرضى الذين أجروا اختبار EIWA قد حصلوا على 24 نقطة من المقياس العام أعلى P < .005) ، (لم يجر الإبلاغ عن أي درجة عن سؤال الحرية)، 25 نقطة من المقياس اللفظي أعلى P < .01) (لم يجر الإبلاغ عن أي درجة عن سؤال الحرية)، و40 نقطة من مقياس الأداء أعلى P < .05)، لم يجر الإبلاغ عن أي درجة عن سؤال الحرية) أكثر من الطلاب الذين قاموا بإجراء WAIS.

اعتمدت الخطة الثانية في دراسة ديفيز وورودريغوز (1979) على استراتيجية مضمون الموضوع. تم اختيار مجموعة من المرضى ثنائيي اللغة (N=12) عشوائياً وتم تعيينهم في واحدة من الإدارات التي تمت فيها موازنة الاختبارات الفرعية للمفردات والاختبارات النفسية العملية في كلٍ من الاختبارين EIWA و WAIS

أجرى اختبار تلك المجموعة ممتحنون ثنائيو اللغة، قدرت محصلات الذكاء اللفظية في اللغة الإنكليزية وقيمة محصلات ذكاء الأداء من مقاييس المفردات ومقاييس الاختبارات النفسية العملية. كانت نتائج تلك الخطة مماثلة إلى حد ما لنتائج البحث الأخير. كانت الدرجات في الاختبار EIWA أعلى 19 درجة من المقياس الكامل عن اختبار WAIS ($P<.05$، لم يسجل أي درجة عن سؤال الحرية)، 22 درجة من مقياس الأداء أعلى $P<.01$) ، لم يجر الإبلاغ عن أي درجة عن سؤال الحرية)، و11 درجة من المقياس اللفظي أعلى (لم يسجل أي درجة عن سؤال الحرية). استنتج الباحثون أن تكافؤ الاختبارين EIWA و WAIS مشكوك به بالرغم من الحجم الصغير للنموذج والمحدودية الواضحة لبعض العوامل الأخرى.

لسوء الحظ، كانت الدراسة السابقة مملوءة بنقاط ضعف خطيرة في المنهجيات مما أدى إلى جعل صدقه في الخارج محدودة. على سبيل المثال: كان المتحنون الذين أجروا الاختبار من بنما ولكن قد جرى تكييف الاختبار وجعل له مقاييس لمجموعات في بورتوريكو. كان كل المتحنين مرضى مصابين بمرض انفصام الشخصية المزمن (80 ٪). تضمن الاختبار 2 من أصل 11 اختبار فرعي، لم تكن هناك أي قياسات موضوعية لثنائية اللغة، وكانت المجموعة صغيرة جداً. مع كل هذا كانت الدراسة مهمة جداً لأنها تسجل المحاولة الأولى المستقلة المطبوعة لتقييم صدق EIWA ومقارنة درجاته مع WAIS.

تقدم الدراسة التي قام بها لوبيز وروميرو (1988) أكثر الفحوصات الشاملة لبنية اختبار EIWA وكان الهدف المحدد لتلك الدراسة تعيين اختلافات محددة بين EIWA و WAIS فيما يتعلق بالإدارة، المحتوى، الدرجات، وصفات معايير النموذج. لكي يشيروا إلى نقاط الاختلافات استخدما WAIS كخط رئيس للمقارنة لأن EIWA استخدم بشكل مباشر. أما فيما يتعلق بتطبيق الاختبار فقد لاحظ لوبيز وروميرو أن هناك خمسة اختبارات فرعية فقط في EIWA متماثلة في الإجراءات التطبيقية مع WAIS أما الباقي فتختلف الأرقام في بداية الاختبار وبعدد الإجابات

الخاطئة قبل التوقف عن إجراء الاختبار الفرعي. أما محتويات الاختبارين فهم يتشاركون في الاختلافات أكثر مما يتشاركون في التماثل. وقد تم اعتبار تجميع الموضوعات في الاختبارات الفرعية (والمدى الرقمي (Digit Span) الوحيدين المتماثلين في الاختبارين وقد وجد الباحثون كثيراً من الاختلافات بخصوص الدرجات. وضع لوبيز وروميرو رسماً بيانياً لتحويل الدرجات الخام (غير منقّحة) إلى مقياس درجات لتلك الاختبارات الثانوية المتماثلة، القياس الرقمي وتجمع الموضوعات. تبين الرسومات البيانية عن ارتفاع ثابت ناتج عن مقياس درجات EIWA لأي مقياس خام. توحي تلك الدراسة أن متوسط الأداء لمعيار نموذج بورتوريكو في اختبار EIWA كان في الغالب أدنى من متوسط أداء نموذج WAIS في الولايات المتحدة. استندت تلك النتيجة على تقدير اختلافات متوسط الدرجات وعلى انحراف درجة القياس عن الوسط؛ لأن المتوسط الحقيقي للدرجات وانحراف درجة القياس للاختبار الثانوي لم يتم نشرها من قبل جمعية علماء النفس. في نهاية الدراسة لاحظ لوبيز وروميرو أيضاً أن معايير النموذج في EIWA تختلف عن تلك في WAIS باختلاف الوضع السكاني مدني/ريفي (P<.001)، على المستوى المهني (P<.001)، والخلفية الثقافية (P<.001) . ناقش الباحثون مدى دلالة نتائجها للممارسين مع التأكيد على أن EIWA يمكن أن ينتج عنه "تضخم درجات" إذا كان الشخص الذي يخضع للاختبار من خلفية ثقافية عالية. اختتم لوبيز وروميرو بحثهما بالقول إنه على باحثي المستقبل استخدام بحث ديفز وورودريغوز (1979) وأن يقوموا بإجراء كل من الاختبارين لمجموعة راشدين عادية.

بحث لوبيز وتوسينغ (1991) عما إذا كان استخدام WAIS-R يمكن أن يقود إلى ضعف تقدير الأداء الوظيفي لناطقي اللغة الإسبانية البالغين وعما إذا كان استخدام EIWA يمكن أن يقود إلى مضاعفة تقدير الأداء لتلك المجموعة. استخدما 47 ناطقاً باللغة الإسبانية و 44 ناطقاً في اللغة الإنكليزية لإجراء اختبار - كان بعض الممتحنين يعانون من مرض الزهايمر- أعطي الممتحنون أربعة اختبارات

فكسلر ثانوية: التماثل، المفردات، المدى الرقمي والاختبار النفسي العملي. قال الباحثون إنه تم اختيار تلك الاختبارات الثانوية لأنها توفر قياسات ذات حساسية لـ neurological impairment ولأنهم متكافئين. ادّعوا أن الاختبارين المدى الرقمي والاختبار النفسي العملي متشابهان تقريباً "الفرق الوحيد بين الاختبارين هو اللغة التي يقام بها الاختبار" (ص 450). في الحقيقة، بالرغم من أن الاختبارين متماثلان فإن بعض البنود في كلٍ من الاختبارين استثنائية. وكان لوبيز وروميرو قد عرفا سابقاً اختبار المدى الرقمي واختبار تجميع الأشياء على أنهما الاختبارين الفرعيين المتماثلين فقط في كلٍ من الاختبارين EIWA و WAIS.

عززت نتائج لوبيز وتوسينغ (1991) نتائج لوبيز وروميرو (1988). وقد أشاروا أن المعايير القياسية لاختبار EIWA تحول الدرجات الخام إلى قيمة قياسية أعلى من معايير القياس لاختبار WAIS-R. استنتج لوبيز وتوسينغ أن EIWA في بعض الحالات يضاعف تقدير الأداء الوظيفي وفي بعض الحالات يعكسها بشكل صحيح. وقد اقترحوا استخدام EIWA مع مجموعات ذات ثقافة محدودة وأحادية اللغة.

درس دمسكي، غاس وغولدن (1997) نموذجين صغيرين في EIWA وذلك بغية إعطائها صدق وجدارة. من الممتع الملاحظة أن كلتا من الدراستين اللتين استخدمتا النسخ القصيرة أو المترجمة التي ليس لهما أي قياسات. اعتمدتا على الاختبارات المترجمة أو المختصرة المعدلة بشكل غير رسمي.

خلاصة بحث EIWA

إن الكتابات عن EIWA قد أظهرت حتى الآن بعض الدراسات التجريبية كما أثبتت أن اختبار EIWA لم يخضع إلى إثبات صدق قابل للقياس مع أهميته واستخدمه التحليلي. عوضاً عن ذلك حقق قبولاً يعود بشكل كبير إلى سمعة اختبار WAIS الذي نشر أكثر من 1.300 دراسة تؤكد ثباته وصدقه.

مقارنة WAIS مع WAIS-R

تقارن هذه الدراسة الدرجات التي حاز عليها الأفراد في EIWA وWAIS-R قد جرى تكييفه في الأصل من WAIS وليس من WAIS-R، يجب ذكر بعض الاختلافات بين WAIS و WAIS-R قارنت كثيراً من المقالات (ليبولد وكليبورن، 1993، اربينا وغولدن وأريل، WAIS 1982) و WAIS-R تشير تلك المقالات مع كتيب WAIS-R إلى أن اختلافات درجات المقياس التامة بين WAISو WAIS-R تكون بقدار 8 درجات على مستوى معامل الذكاء. في كل دراسة كانت درجات WAIS أعلى من درجات WAIS-R.

أكّدت التغييرات بين درجات WAIS و WAIS-R على الحاجة للانتباه إلى موضوعات ثلاثة بالتحديد. أولاً: لكون كل الدرجات قد تم الحصول عليها من مجموعة معيارية فيجب أن تكون المجموعة الممتحنة مناسبة. ثانياً: لأن المجموعة الممتحنة تتغير مع الوقت فإن بعض البنود في الاختبار قد تصبح قديمة بعض الشيء؛ ولذلك يجب مراجعة الاختبارات وإعادة معايرتها من وقت لآخر. ثالثاً: من الضروري إعادة تكافؤ الاختبارات التي جرى تعديلها أو تحديثها مع الاختبارات السابقة. عندما يجرى تكافؤ الاختبارات فإن ذلك يجعل اتخاذ قرارات حسب مقياس معتمد (معياري)، ودون عمليات التكافؤ فإن القرارات التحليلية المعتمدة على معامل الذكاء سوف تظهر اختلافات مع الوقت وعبر نسخ الاختبارات المختلفة.

المنهجية والإجراءات:

تستخدم الدراسة الموجودة مجموعة من 50 رجلاً وامرأة من بورتوريكو أعمارهم بين 17 و 59 سنة، وهم إما من مواليد بورتوريكو أو أن آباءهم البيولوجيين من مواليد بورتوريكو. تم استبعاد المرشحين الذين أظهروا أعراضاً ذات دلالة على مرض عقلي. سجلت الخلفية الثقافية والمهنية للممتحنين ولكنها لم تستخدم كمعيار لتضمينها أو استبعادها من الدراسة.

إن المعايير الرئيسة في اختيار الممتحنين كانت في ميدان المهارة اللغوية: لذلك يستخدم التحليل الحالي خطة في نظام الموضوع (within-subject) وكان أساسياً أن الاختلافات اللغوية تساوي إلى أقصى حد ممكن. لهذا، لكي يكون الممتحنون ضمن الدراسة يجب عليهم الحصول على مستوى الدرجات ذاتها في اختبار المهارة في كلٍ من اللغتين الإنكليزية والإسبانية المقاسة حسب المقياس الثاني لقواعد النحو ثنائي اللغة Bilingual Syntax Measure II بيرت، دولي، هيرناندز، تاليبوروز، (1980).

إذا كان هناك اختلاف كبير للممتحنين بالمهارة في اللغة الإنكليزية والإسبانية عندئذ فإن أي اختلافات من الممكن وجودها في معامل الذكاء التام تعزى إلى الاختلافات في الطلاقة اللغوية للممتحنين. لذلك كان التأكيد على تساوي المهارة اللغوية للممتحنين مهماً. إذا لم يجر استخدام التحليل الذي سبقت الإشارة إليه فإن خطة العمل التي تقيّم بدقة أي اختلافات بين EIWA و WAIS-R سوف تكون خاطئة، لأن تلك الاختلافات قد تعكس إلى حد ما الاختلافات اللغوية.

توجد قلّة من الاختبارات التي تقوِّم هيمنة اللغة بشكل فعال. وجد أوكلاند، دلونا مورغن (1997)، عند مراجعتهم 27 قياساً، أن أربعة فقط تقدم معلومات عن كلٍ من الصدق والثبات. إن نقص التقصي الكافي لتلك الأدوات توجب علينا الحذر عند استعمالها. ثلاثة معايير مهمة قادتنا إلى اختيار هذا الاختبار للمهارة اللغوية لهذه الدراسة. الأول: هو أن الاختبار بحاجة إلى دعم من بحث يُصدّق على صدقه وثباته. الثاني: هو أنه يحتوي على سلسلة من القياس تتضمن على الأقل التطور اللغوي في المستوى الأعلى المدرسي، وأخيراً لتجنب التشابك مع مقاييس الذكاء كان الاختبار المناسب الأفضل هو المقياس الثاني لقواعد النحو ثنائي اللغة الذي جرى ذكره سابقاً.

إدارة الاختبار:

خضع كل المتحنين في هذه التجربة إلى الإجراءات نفسها بشكل تام. طُلب من المتقدمين الذين تم اختيارهم حسب المقاييس السابقة التوقيع على استمارة موافقة على إجراء الاختبار كما تم إعطاؤهم وصفاً مختصراً للدراسة. أُخبر الممتحنون أن الهدف من الدراسة هو جمع معطيات تساعد على اكتشاف مقارنةEIWA و WAIS-R كما تمت الإجابة على الأشخاص الذين طلبوا معلومات مفصّلة عن الافتراض العلمي في الدراسة أن تلك المعلومات سوف تعطى لاحقاً بعد إتمام الاختبار مع إمكانية استرداد ورقة اختبارهم من الدراسة في حال رغبتهم بذلك. لم يسترد أحد من الممتحنين ورقة الاختبار بعد الانتهاء من إجراء الاختبار. بعد تلك المقدمة المختصرة أعطي الطلاب اختبار المقياس الثاني لقواعد النحو ثنائي اللغة (BSM II). عملياً تم الافتراض بشكل مبدئي أن كل الممتحنين قد تم قبولهم للاشتراك في هذه الدراسة لذلك تم اختبارهم باستخدام EIWA أو WAIS-R تم استبعاد ورقة اختبار لاحقاً لأن قياس اختبار الممتحنين أظهر اختلافاً في مهارته اللغوية. كل شخص اشترك في هذه الدراسة سجل درجات في فئة المهارة اللغوية في اللغة الإنكليزية والإسبانية في اختبار BSM II.

أُعطي الطلاب الاختباران EIWA أو WAIS-R في تسلسل متوازن كي يتم السيطرة على أثر التنظيم: التقليل من أثر التدريب (المران) ثم القيام بكل اختبار على حدة، وجرى تقييم الدرجات من قبل عالم نفس ثنائي اللغة حسب مقياس الإرشادات لكل اختبار.

النتائج:

جرى اختبار 50 فرداً بين عام 1984 - 1992 وتم الحصول على كل الأفراد من مؤسسة الصحة العقلية العامة والمؤسسات التربوية في نيويورك ومنها منطقة وستشستر، ضاحية في شمال نيويورك. كان هناك 31 أنثى و 19 ذكراً في تلك

المجموعة المكونة من 50 فرداً وتراوحت أعمارهم بين 17 و 59 عاماً بمتوسط 31.9 وانحراف قياس 8.77 تراوحت عدد سنين الدراسة بين 4 و 21 بمتوسط 1.14 وانحراف قياس 3.05. يظهر الجدول 9-1 النتائج الإجمالية التي تم الحصول عليها للممتحنين في الاختبار اللغوي واختبار الأداء ومقياس الدرجات التام في كلٍ من EIWA أو WAIS-R.

الجدول 1.9

وصف إحصائي لدرجات الاختبار للطلاب

	EIWA			*WAIS-R*		
	Verbal	*Performance*	*Full Scale*	*Verbal*	*Performance*	*Full Scale*
M	116.6	125.4	121.5	92.4	97.4	93.8
SD	10.52	9.63	9.23	11.85	12.37	11.44
Maximum	137	144	140	118	130	126
Minimum	88	106	97	69	75	73

تم إجراء تحليل Hotelling T^2 لاختبار عما إذا كانت هناك أي اختلافات بين الاختبارات الثانوية الإحدى عشر. كانت إحصائيات Hotelling - Lawley Trace statistic مهمة على مستوى أعلى من 001 (ويلكس 1=0.390 df (df درجة الحرية) = 40) لأن Hotelling T^2 كانت مهمة، أجريت عملية ملاءمة 14 زوجاً من اختبارات t في الاختبارات الثانوية الإحدى عشرة والدرجات الثلاثة المجموعة. تظهر نتائج اختبارات t في الجدول 9-2 .

إن النتائج التي تم الحصول عليها في الجدول 9-2 تظهر وجود اختلافات واضحة ومهمة (P< 0.001) ، (df=49) بين المقياس التام، المهارة اللغوية، وأداء معامل الذكاء في كلٍ من EIWA وWAIS-R أظهرت دراسات أخرى للاختبارات الثانوية أن تلك الاختلافات ليست نتيجة خلل في واحد أو أكثر من الاختبارات، ولكن نتائج الاختبارات تقول إن تلك الاختلافات تعود إلى اختلافات الاختبارات الثانوية (P<.001)، (df =49)

جرى إقامة تحليل خاص على الاختبارات في الثانوية اللغوية لإقامة الدليل على مساواة WAIS-Rو EIWA.

الجدول 2.9

نتائج متماثلة لأزواج الاختبار T

IQ or Subtest	*WAIS-R Mean*	*WAIS-R SD*	*EIWA Mean*	*EIWA SD*	*Mean Difference*	*t*
Full Scale IQ	93.8	11.4	121.5	9.2	27.7	33.7*
Verbal IQ	92.4	11.9	116.6	10.5	24.2	26.3*
Performance IQ	97.4	12.7	125.4	9.6	28.0	25.2*
Information	8.6	2.4	13.1	2.6	4.5	16.1*
Digit Span	8.4	2.7	12.5	2.8	4.1	11.5*
Vocabulary	9.3	2.7	12.1	2.2	2.8	8.4*
Arithmetic	8.1	2.3	12.6	2.5	4.5	19.5*
Comprehension	9.0	2.6	14.1	2.7	5.1	15.8*
Similarities	9.0	2.8	13.5	1.6	4.5	14.5*
Picture Completion	9.0	2.7	14.1	1.9	5.1	15.7*
Picture Arrangement	9.5	3.0	13.7	2.1	4.2	10.7*
Block Design	9.0	2.4	13.6	2.0	4.6	18.6*
Object Assembly	9.3	2.6	14.8	2.4	5.5	18.6*
Digit Symbol	9.5	2.7	14.2	2.6	4.7	22.6*

*(P< .001, df =49)

تم إخضاع ذلك الاختبار الثانوي إلى تحليلات منفصلة لاكتشاف عمّا إذا كانت عملية ترتيب البنود وتسلسلها يتوافق مع مستوى صعوبتها التي يعاني منها المتحنون. كان ترابط البنود الهرمي في WAIS-R + 934. ، وهذا يعني التوافق بين طريقة إجابة المتحنين للبنود وتسلسلها. على العكس من ذلك كان الترابط في 777+EIWA. ، أي انخفاض واضح (F = 3.158; P< .001)؛ في القيمة المطلقة للترابط. ترابطت الاختبارات الثانوية للمفردات بنسبة 80 (P< .001,df = 47)، مع المقياس التام لعامل الذكاء. كانت الدرجات المساوية في EIWA 66 (P< .001, df = 47) كان الفرق بين الدرجات المساوية ليس ذا دلالة إحصائية تظهر كل الدرجات المتعلقة بالاختبارات الثانوية في الجدول 9-3.

الجدول 3.9

معاملات الارتباط بين الاختبارات الثانوية والمقياس التام لعامل الذكاء

Subtest	*WAIS–R Full Scale IQ*	*EIWA Full Scale IQ*	*df*
Vocabulary	.80	.66	47
Comprehension	.77	.78	47
Arithmetic	.72	.73	47
Similarities	.71	.64	47
Information	.71	.63	47
Digit Span	.68	.60	47
Picture Arrangement	.64	.66	47
Block Design	.62	.61	47
Object Assembly	.54	.61	47
Picture Completion	.52	.73	47
Digit Symbol	.48	.51	47

المناقشة والمدلولات:

أظهرت هذه الدراسة أن هناك فروقاً ذات دلالة في الدرجات التي حصلت عليها المجموعة الثنائية اللغة عند إجرائها كل من الاختبارين EIWA و WAIS-R. أظهر المقياس التام لمعامل الذكاء متوسط اختلاف 27.7 درجة. كما أظهر كل من اختبار المهارة اللغوية واختبار أداء معامل الذكاء اختلاف درجات 24.2 و 28.0 على التوالي. كانت تلك الفروقات ذات دلالة إحصائية عالية وذات معنى تطبيقي.

إن الاختلاف بمقدار 27.7 (انحراف درجتان تقريباً) بين درجات المعايير التامة في الاختبارين لها تفسيرات مروعة للمختصين الإكلنيكيين. بالطبع إن المرء يستطيع أن يفكر بشكل عقلاني أن كلاً من الدرجات صحيحة-إن EIWA يعطي درجات الشخص حسب معيار نموذج 1967 وإن WAIS-R يعطي درجات حسب معيار نموذج 1980 ولكن ذلك الموقف العقلاني ليس مفيداً للأغراض العملية التي أجريت الاختبارات بسببها. يجب ملاحظة أن ذلك الفرق في الدرجات يمكن أن يكون 8 درجات أقل من المعيار التام لو استخدم WAIS عوضاً عن WAIS-R في ذلك التحليل. على كل فإن الاختلاف أكبر من انحراف معيار تام وذي نتائج مهمة.

قبل الحصول على الدلائل التجريبية لتفاوت الدرجات. انطلقت بعض الدراسات في محاولة لشرح الاختلافات المفترضة. اقترحت إحدى النظريات أن الدرجات العالية في اختبار EIWA يمكن أن تكون انعكاساً للحساسية الثقافية للاختبار. تلمح هذه الفكرة إلى أنه لكون الاختبار قد جرى تهيئته في بورتوريكو فإن محتوى البند أو على الأقل قسم المفردات في الاختبار لها مدلول أكبر للمجموعات في بورتوريكو. وبمتابعة هذه الفكرة فإنه من المتوقع أن يكون معامل الذكاء اللغوي يعكس تفاوتاً أكثر في الدرجات؛ لأن اللغويات كما هو معروف ذات غنى ثقافي كبير.

تظهر نتائج هذه الدراسة أن اختلافات معامل الذكاء في الأداء كانت أكثر من معامل الذكاء اللغوي؛ لذلك فإن القول بأن الدرجات الأعلى للاختبار تعكس

الحساسية الثقافية لم يجر تأييده بتلك النتائج إلا إذا رغب الباحث أن يؤيد الفكرة القائلة إن الاختلافات الثقافية يمكن أن تنعكس في المعيار الثانوي للأداء إلى حد كبير من المعيار اللغوي. بالطبع، إن الثقافة تؤثر على الأداء بالإضافة إلى تأثيرها على ميدان الذكاء اللغوي، ولكن الاختلافات في الاتجاه المعاكس محتملة.

أما الفكرة الثانية وهي أن الاختلافات تنشأ لأن الاختبارين يقيسان بشكل أساسي سمات نفسية مختلفة، لم تؤيد هذه الدراسة الفكرة لعدم وجود اختلافات ذات مدلول في الطريقة التي يرتبط فيها EIWA و WAIS-R مع درجات المقياس التام الخاص بكل اختبار. وأكثر من ذلك فإن دراسة تحليل العوامل لبنية اختبار EIWA غوميز وآخرون، 1992 تثبت أن بنية الاختبار كانت انعكاساً تاماً لاختبار WAIS.

يوحي الدليل الأكثر إقناعاً أن هناك مشكلات مستمرة في اختبار EIWA، وقد ظهر سابقاً أن نموذج المعيار يختلف عن القيمة الأصلية للمجتمع الإحصائي حسب المنطقة، المهنة، العمر والثقافة. على كل حال، فإن الدليل المقام يوحي بأن الاختلافات أكثر تعقيداً. وللتحقق من أن هناك صعوبات في مستوى الاختبار الثانوي، جرى التدقيق في الاختبار الثانوي اللغوي.

تم اكتشاف أن العلاقة بين صعوبة البند وبين مكان ترتيبه في الاختبار الثانوي للمفردات كانت أقل في EIWA منها في WAIS-R يعطي هذا التباين مصداقية أكثر لشكوى الإكلنيكيين من أنهم لم يصلوا فقط إلى "اختبار إلى الحد الأقصى" في اختبار EIWA الثانوي للمفردات. يقرر علماء النفس غالباً أنه يجب إجراء الاختبار الثانوي الكامل لأن الممتحن يسجل درجات لكلمات صحيحة لعدة مرات بشكل متقطع قبل تسجيله سبعة أخطاء متتالية الذي بدوره يؤدي إلى عدم استمرار الاختبار. (في اختبار WAIS خمسة أخطاء متتالية تنهي استمرار الاختبار) لا تضيف تلك التظاهرات عدم المصداقية إلى الاختبار فقط لكنها تزيد من إحباطات

الممتحنين، الذي يحتمل الوضع المخزي بمواجهة عدد كبير من المفردات التي يجهل معناها.

هناك الكثير من الملاحظات المثيرة للانتباه المتعلقة بالدراسة الحالية وبدراسات غرين ومارتينز (1967) اللذين قدما EIWA بالرغم من أن اختبار EIWA يستخدم بشكل رئيس فإنه عانى من مشكلات جديّة. تتضمن هذه المشكلات على الأقل الشك في قياس الأداة وموضوع تفسير الدرجات المتعلق به. إن الموضوعات التي أثارها ميليندز (1994) لوبيز وروميرو (1988) تعيّن أسباب حدوث تضخم الدرجات: إن السلوك في اختبار EIWA يؤدي إلى درجات أعلى مما يؤديه السلوك ذاته في اختبار WAIS وإن الأساس المنطقي لهذا الحدث ببساطة هو أن متوسط الأداء الذي يجري اختباره في نموذج قياس EIWA كان أدنى من ذلك في WAIS تلك الاختلافات ناشئة عن الاختلافات الإجمالية في الخلفيات الثقافية والحالة الاجتماعية والاقتصادية بين الولايات المتحدة وبورتوريكو. عندما وضع متوسط الأداء في EIWA 100 أصبح متوسط الفرق في الدرجات جوهرياً في الاختبار.

توثّق دراسات ميليندز أن درجات اختبار EIWA لا تعني ذات الشيء كدرجات اختبار WAIS من حيث المهارة المعرفية المتوقعة. هناك حاجة ملحّة لإعادة معايرة EIWA. أشار غرين ومارتينيز (1967) أن "هناك حاجة لإعادة المعايرة في الولايات المتحدة كل 12 - 18 عاماً" يستند تعليقهم هذا على الخبرة المكثّفة مع WAIS في الولايات المتحدة. أكدوا على أن إعادة المعايرة في منطقة مثل بورتوريكو، حيث تتغير صفات السكان بشكل متسارع، ضروري جداً.

بالرغم من الإنذار المبكر الذي أطلقه غرين ومارتينيز (1967) فإنه لم يجر أي تعديل لاختبار EIWA من قبل هيئة علماء النفس منذ أن تم نشر الاختبار الأساسي عام 1968. بصرف النظر عن الموضوعات الرئيسة للحفاظ على اختبار ذي مقياس سيكولوجي سائد فإن كتيب اختبار EIWA موبوء بأخطاء صغيرة مزعجة كان يجب

إصلاحها. أحد هذه الأخطاء هو أن الكلمات المطبوعة في الكتيب بينها فراغات شاذّة. على سبيل المثال، البند الأول في الاختبار الثانوي للمعلومات يستخدم الكلمة الإسبانية Pajaros ولكنها مطبوعة "Paj aros"، يستخدم البند الثاني كلمة Pelota ولكنها طبعت "Pel ota"، ويستخدم البند الثالث كلمة yrba وهي مطبوعة "yer ba" أما البند الرابع فيستخدم كلمة Planta وهي مطبوعة "Pl ant a" إن بنود الاختبار ليست المكان الوحيد التي وقعت فيها أخطاء. في المقطع الأول فقط (خمس أسطر) المتعلق بالإرشادات حول الاختبار الثانوي للمعلومات هناك على الأقل عشرة أخطاء في تحديد ترتيب الكلمات. هذا النوع من الخطأ - عندما تجتاح الفراغات العشوائية تكامل الكلمات - منتشر في الكتيب وهو مزعج جداً لأي ناطق باللغة الإسبانية.

بالرغم من معرفة المرء باللغة، فإن الفراغات ضمن الكلمات تشتت الانتباه وتسبب توقفاً مؤقتاً أثناء قراءة بنود الاختبار. حتى إن علامات النطق في الكتيب تبدو وكأنها مخططة بالقلم. ما يثير الانتباه هو أن أخطاء الفراغات في الطباعة لم تكن موجودة في أقسام النسخة النهائية التي قدمها غرين ومارتينيز (1967) ولكنها موجودة في الكتيب المطبوع. إن تأثيرات تلك الأغلاط المشتتة للانتباه في المواقف الاختبارية ليست معروفة بالتحديد ولكنها دون شك تعرقل التقويم الدقيق.

إن قياسات الجودة ليست السمة المميزة لاختبار (EIWA) الذي تشرف عليه هيئة علماء النفس، ويظهر ذلك بالفروق الواضحة بينه وبين الشكل المصقول الموجود في كتيب اختبارات WAIS-R و WAIS-III. على كل فإن EIWA لا يلائم القياسات المحترفة المعاصرة.

يجد بعض الباحثين الميدانيين بالمقارنة مع WAIS-R أن اختبار EIWA "سهل جداً" وأنه يعطي درجات "متضخمة". ما يدعو للدهشة أن غرين ومارتينيز شكوا من "أن مستوى صعوبة الاختبار عال جداً" (ص10)، وإن إصلاح ذلك يتطلب "عاماً أو عامين من الجهد في مجال تطوير البند ومجال الاختبارات"(ص. 52). كيف يمكن

أن يكون هناك ذلك التباين؟ إن هذه الملاحظات المتعارضة يمكن أن تكون نتيجة نموذج معايرة سيئ، أو يمكن أن يكون انعكاساً للتغيرات السريعة في صفات السكان. لا يمثل نموذج المعايرة مع مضي الوقت السكان الذين يتطورون بسرعة؛ لذلك يجب التأكيد على إعادة المعايرة. تؤكد الدراسة الحالية على التباين الواضح بين EIWA و WAIS-R وعلى مقدرة الاختبار على مجاراة المعايير لمجموعة تتغير بسرعة في مجال الثقافة في القيمة الأصلية للمجتمع.

يحتكر اختبار EIWA افتراضياً في الوقت الحالي اختبار الذكاء للبالغين ذوي الأصول اللاتينية في الولايات المتحدة. إن الأعداد المتزايدة للمجموعات من أصول لاتينية وازدياد حاجاتهم سوف تؤدي عاجلاً أم آجلاً إلى الحاجة إلى قياس/ معيار المقدرة التنموية في اللغة الإسبانية. وقد تم البرهان على ذلك باتفاق الآراء في الدراسات الحديثة على أن معايير EIWA ناقصة وقديمة وأن الحل الوحيد هو في إعادة عملية التكييف والقياس لهذا الاختبار. إذا تمت إعادة صياغة هذا الاختبار فهناك أمل أن يستطيع تطبيق المقياس الحالي لتكييف الاختبار (كيسنجر، 1994، هامبلتون، (1994).

يظهر هذا البحث أن هناك فروقات كبيرة بين الاختبارين. ظهر أن EIWA فيه مشكلات كثيرة تبدأ في موضوعات بناء البند حتى موضوعات المعايير التي تسبب كارثة لأي اختبار يعتمد على معايير سكانية. ظهر على مستوى بنية الاختبار أن هناك صعوبة في الحصول على نموذج قياس يطابق متغيرات الإحصاء السكاني. كما ظهر أيضاً أن الاختبار الثانوي للمفردات في EIWA يختلف بشكل كبير عن اختبار WAIS-R في دقة ترتيب البنود حسب درجة الصعوبة. توحي هذه النتائج مع نتائج ميليندز (1994) أن إجراءات بناء اختبار EIWA التي جرت عام 1967 -1968 قد تكون ناقصة/ خاطئة، أو لو كانت جيدة في ذلك الوقت فهي قديمة جداً في الوقت الحالي، وإن المجموعات السكانية ذات الأصول اللاتينية في التسعينيات

مختلفة بشكل كبير عن سكان بورتوريكو في 1967. يبقى اختبار EIWA محاولة أولية أساسية لإحداث تكييف WAIS إلى اللغة الإسبانية. لسوء الحظ، فقد تم إغفاله من قبل هيئة الاختبارات بعد البدء باستخدامه ولم تجر الاستفادة من تدقيقه أو تحديثه كما جرى لمثيله WAIS.

توحي الدراسة الحديثة بشكل قوي إلى أن EIWA لم يعمر بشكل جيد ولذلك يجب إعادة قياساته، والأكثر من هذا فإن نتائج هذه الدراسة تظهر أنEIWA وWAIS-R ليسوا متكافئين وظيفياً. في ضوء هذه الدراسة فإن استمرار استخدام علماء النفس لهذا الاختبار دون تعديل جدِّي لتفسير الدرجات يثير أسئلة أخلاقية خطيرة.

شكر

إن هذا الفصل مبني على رسالة بحث بإشراف قسم علم النفس في جامعة فوردهام أجراه مالدونالدو بإشراف كيسنجر. يدين الكتاب بالشكر للدكتورة جانيت ف. كارلسون لقراءتها النقدية لهذا البحث. بالطبع أي أخطاء هي حاصلة من الباحثين.

*****

المراجع

Burt, M. K., Dulay, H. C., Hernandez, E., & Taleporos, E. (1980). *Bilingual Syntax Measure II technical handbook*. New York: Psychological Corporation.

Conde López, V., & Domeneca López, B. (1977). Algunas reflexiones sobre las adaptaciones españolas de las Escalas de Wechsler para Adultos [Some reflections about the Spanish adaptation of the Wechsler Scale for Adults]. *Revista de Psicología General y Aplicada, 32*, 619–645.

Davis, T. M., & Rodriguez, V. L. (1979). Comparison of WAIS and EIWA scores in an institutionalized Latin American psychiatric population. *Journal of Consulting and Clinical Psychology, 47*, 181–182.

Demsky, Y. I., Gass, C. S., & Golden, C. J. (1997). Common short forms of the Spanish Wechsler Adult Intelligence Scale. *Perceptual & Motor Skills, 85*, 1121–1122.

Demsky, Y. I., Mittenberg, W., Quintar, B., Katell, A. D., & Golden, C. J. (1998). Bias in the use of standard American norms with Spanish translations of the Wechsler Memory Scale–Revised. *Assessment, 5*, 115–121.

Eyde, L. D. (1992). Introduction to the testing of Hispanics in industry and research. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 167–172). Washington, DC: American Psychological Association.

Geisinger, K. F. (Ed.). (1992). *Psychological testing of Hispanics*. Washington, DC: American Psychological Association.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.

Gómez, F. C., Piedmont, R. L., & Fleming, M. Z. (1992). Factor analysis of the Spanish version of the WAIS: The Escala de Inteligencia Wechsler para Adultos (EIWA). *Psychological Assessment, 4*, 317–321.

Green, R. F. (1964). Desarrollo y estandarización de una escala individual de inteligencia para adultos en español [Design and standardization of the individual intelligence scale for adults in Spanish]. *Revista Mexicana de Psicología, 1*(3), 231–244.

Green, R. F., & Martinez, J. (1967). *Standardization of a Spanish-language adult intelligence scale* (Final Report, Project No. 1963, Contract No. O. E. 3-10-128). Washington, DC: U.S. Department of Health, Education and Welfare.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58–79). Washington, DC: National Academy Press.

Herrans, L. L. (1969). Sex differences in the Spanish WAIS scores. *Dissertation Abstracts International, 30*, 1432–1433.

Herrans, L. L. (1973). Cultural factors in the standardization of the Spanish WAIS or EIWA and the assessment of Spanish-speaking children. *School Psychologist, 28*, 27–34.

Lippold, S., & Claiborn, J. M. (1983). Comparison of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale–Revised. *Journal of Consulting and Clinical Psychology, 51*, 315.

López, S. R., & Romero, A. (1988). Assessing the intellectual functioning of Spanish-speaking adults: Comparison of the EIWA and the WAIS. *Professional Psychology: Research and Practice, 19*(3), 263–270.

López, S. R., & Taussig, I. M. (1991). Cognitive-intellectual functioning of Spanish-speaking impaired and nonimpaired elderly: Implications for culturally sensitive assessment. ***Psychological Assessment, 3,*** 448–454.

Macias, R. F. (1977). Hispanics in 2000 A.D.—projecting the number. ***Agenda,*** 7(3), 16–20.

Martinez-Urrutia, A., & Spielberger, C. D. (1973). The relationship between state–trait anxiety and intelligence in Puerto Rican psychiatric patients. ***Revista Interamericana de Psicologia, 7,*** 199–214.

McShane, D., & Cook, V. J. (1985). Transcultural intelligence assessment: Performance by Hispanics on the Wechsler scales. In B. B. Wolman (Ed.), ***Handbook of intelligence: Theories, measurements, and applications*** (pp. 737–785). New York: Wiley.

Melendez, F. (1994). The Spanish version of the WAIS: Some ethical considerations. ***Clinical Neuropsychologist, 8,*** 388–393.

Oakland, T., DeLuna, C., & Morgan, C. (1977). Annotated bibliography of language dominance measures. In T. Oakland (Ed.), ***Psychological and educational assessment of minority children.*** New York: Brunner/Mazel.

Urbina, S. P., Golden, C. J., & Ariel, R. N. (1982). WAIS/WAIS–R: Initial comparisons. ***Clinical Neuropsychology, 4,*** 145–146.

Wechsler, D. (1968). ***Manual para la Escala de Inteligencia Wechsler para Adultos*** [Manual for the Wechsler Intelligence Scale for Adults]. New York: Psychological Corporation.

10

تطوير الاختبارات للاستعمال في اللغات والثقافات المتعددة: التماس للتطوير المتزامن

نوربرت ك. تانزر
جامعة أليانت العالمية

إن عولمة السوقِ الاقتصادية، وقابلية الحركة المتزايدة لقوة العالم العاملة، وظهور المجتمعاتِ المتعدّدة الثقافاتِ المعقدةِ، مثل الإتحاد الأوربي، قد جَلَبَ تحديات جديدةَ إلى الاختبار التربويِ والنفسيِ الذي تجاوز المجالَ التقليدي للتقييمات الأحادية اللغة/ الأحادية الثقافة. مع الوقت، أصبحت الحاجة لأدوات صالحة في التقييمات المتعددة الثقافة/ المتعددة اللغة، أكثر أهمية (بارترام وكوين، 1998، هو وأوكلاند، 1991، أوكلاند، 1997، 2004، أوكلاند وهو، 1992).

لقد تم القبول الآن على نحو واسع، بأن تطور أدوات مثل تقييمات أو اختبارات متعددة الثقافة/ متعددة اللغة تستلزم أكثر من مجرد الترجمة، أي إعادة كتابة نص ما من لغة إلى أخرى. على الرغم من أن فريقاً من المترجمين المحترفين يستخدمون طريقة الترجمة - ترجمة مرتجعة (بريسلين، 1980)، قد ينتجون صيغ اختبارات متعددة اللغات متكافئة لغوياً. هذه النسخ لا تَشترك في المعنى النفسي نفسه بالضرورة.

إن الترجمة الصحيحة للمادة "شاهدت التلفاز أكثرَ من المعتاد" ستَبْقى مادةُ غير ملائمةُ عندما تُقدّمُ إلى سكان الساحل الذين لا يوجد لديهم كهرباءُ في بيوتِهم. (فان هامتن وفان دي فيفر، 1996).

يظهر هذا المثال بشكل واضح أن التقييمات المتعددة الثقافة/ المتعددة اللغة تَتطلّبُ معالجة متوازنة للاعتبارات النفسية، والقياسية، واللغوية والثقافة المتقاطعة، والثقافة، وذلك لكي تَضْمنَ صدق التركيب (تكافؤ التركيب). صمّمَ الاختبار لقيَاْسه ("تكافؤ أداة")، إدارة الاختبار (تكافؤ الإدارةِ")، وقد استخلصت بعض النتائج من علامات الاختبار (براكين وبارونا، 1994، بريسلين، 1980، 1986، غيسينجير، 1994، غرينفيلد، 1997، هامبلتون، 1994، فان دي فيفر ولونغ، 1997a، 1997b، فان دي فيفر وتانزر، 1997). لاحظ أن الجزء المهم لعملية التثقيف الثقافي (أو الاجتماعي) يُنقل عبر اللغة. وهكذا، فإن الأشخاص الذين لديهم لغات أمّ مختلفة لديهم أيضاً خلفيات ثقافية مختلفة. لهذا، فإن التقييمات التي تُجرى بصيغ لغوية مختلفة للممتحنين الذين لديهم لغات أمّ مختلفة هي أيضاً على الأغلب متعددة الثقافات. بناءً على هذا، التقييمات المتعددة اللغات عادة ما تكون تقييمات متعددة الثقافات. لاحظْ أيضاً بأنّ سكان المجموعات العرقية المختلفة، حتى لو كانوا يَتكلّمونَ لغة مشتركة، (أي: أغلبية/ أقلّيات)، (ما زالَتْ تُشكّلُ مجموعات ثقافية مختلفةً. من هنا، فإن التقييمات في المجتمعات الأحادية اللغة، المتعددة الأعراق، تكون متعددة الثقافات، حتى لو أُجريت بالصيغة اللغوية نفسها لاختبار ما.(مثال، "التقييمات المتعددة الثقافات/ الأحادية اللغة").

تكافؤ البنية

بما أن الاختبارات الأحادية الثقافة/ الأحادية اللغة يمكن أن تطور بالاستناد على النظرياتِ السائدةِ للتراكيبِ التي يَنْوونَ قيَاْسها، فإن تطوير أدوات صادقة للتقييمات المتعددة الثقافة والمتعددة اللغة تَتطلّبُ قابلية التعميم للتقاطع الثقافي للتراكيب عبر كل الثقافات واللغات المقصودة. في تكوين فكرتهم عن التراكيب

النفسية، يميل مطورو الاختبارات الأحادية الثقافة إلى التعرض لخطر التحيز العرقي مما قَدْ يُهدّدُ صدق الاختبارات عبر الثقافات. على سبيل المثال، بينت دراسات سابقة متراكمة أن أدوات قياس الشخصية الغربية لا تحدد هيئة الشخصية الصينية لأحد السكان الأصليين مثل "الوجهِ" و"الانسجام" (سونغ، 2004، شونغ وآلو 1996، يانغ وبوند، 1990، زانغ وبوند، 1998).

تكافؤ الإدارةَ والأداة

علاوة على، وما وراء، تكافؤ التراكيب المتقاطعة ثقافياً ولغوياً، يجب على مطوري الاختبارات أن يقدموا دليلاً على تكافؤ التقاطع الثقافي واللغوي لأدواتهم (أي: غياب "تحيّز الأداة") والإجراءات المتعلقة بإدارتهم (أي: غياب "تحيّز الإدارة"). ومن أجل تسهيل هذه المهمّةِ، قامت لجنة الاختبار الدوليةِ (ITC) بتحضير اختبارات دلائل التكييف ITC، وهي مجموعة تتألف من 22 دليلاً للممارسات الموصى بها لتطوير الاختبارات المتعددة الثقافة/ المتعددة اللغة، التي ذُكرت في الفصل الأول من هذا الكتاب (انظر أيضاً الى هامبلتون، 1994، 2002، فان دي فيفر هامبلتون، 1996). لتَأكيد سياقِ التقييمات المتعددة اللغاتِ والمتعدّدة الثقافاتِ، يقوم الدليل بتوجيه عدة معايير للتطوير (توجيهات 3-12)، الإدارة (توجيهات 13-18)، التوثيق (توجيه 19) لأدوات التقاطع الثقافي واللغوي، بالإضافة إلى تفسيرِ علامات الاختبار (توجيهات 20-22).

إن دلائل التكييف ITC تؤكّدُ، بشكل خاص، على الحاجة الى تقديم بينة على صدق التقاطع الثقافي للتركيب (توجيه رقم 2) والاختبار (توجيهات 1، 3-6، 10-11) عبر كُلّ السكان واللغات المقصودة. وقد أكدوا أيضاً على أهمية تَحديد وتَبرير مقارناتٍ صادقة بين النَتائجِ (مثال، علامات الاختبار) التي تم تحصيلها من الصيغ اللغوية المختلفة أو مِنْ المجموعاتِ الثقافيةِ المختلفةِ (توجيهات 12، 14، 20-22). علاوة على ذلك، يُؤكّدونَ أيضاً على تَوثيق (توجيهات 4، 17، 19) كُلّ التغييرات التي

كَانتْ ضرورية أثناء عملية التكيّفِ أو التطويرِ التي تضمنت محتويات السؤال أو مادّةِ المحفّزِ (توجيهات 6، 14)، صيغ إجابة (توجيهات 5، 14)، تعليمات الاختبار (توجيهات 4، 16)، إجراءات الإدارةِ (توجيهات 5، 15، 16، 18)، وقواعد العلامات (توجيهات 12، 20-21). أخيراً، يُؤكّدونَ على أهمية تطبيق تقنيات إحصائيةَ وتصاميمَ مناسبة ومجموعة بيانات لاكتشاف المصادرِ المحتملة لعدم التكافؤ ولدِراسةَ فعالية أيّ قياسات مقابلة تم أخذها (توجيهات 7-11، 13). أخيراً وليس آخراً، إن مؤهلات مستخدمي الاختبارات لإجراء تقييمات متعددة الثقافة/ متعددة اللغة (توجيهات 14، 18) هي أيضاً قد تم إلقاء الضوء عليها.

التطورات المتزامنة مقابل التطورات المتعاقبة لأدوات متعلقة بحالاتِ التقييمِ المتعدّدة الثقافاتِ والمتعددة اللغاتِ

بالرغم من أن السبب الجوهري وراء كُلّ توجيه تم التعليق عليه وتوضيحه بالأمثلةِ التجريبيةِ (فان دي فيفر وهامبلتون، 1996)، فإن دلائل تكييف الاختبار (ITC) معيارية أكثر منها إرشادية. فهي لا تقدم إستراتيجيات للحُصُول على اختبارات صادقة متعددة الثقافات/ متعددة اللغات. ولا تناقش حسنات وسيئات إيجاد طرق معالجة.

في طريقة معالجة تطور اختبار المحاكاة، تم تطوير أداة جديدة للاستعمال في عدد منْ مجموعاتِ ثقافيةٍ محددة مسبقاً (ثقافات مرجعية) و/ أو لغات (لغات مرجعية). يستلزم ذلك عادة " طريقة معالجة لجنة"، أي إنها قدرة عمل متعددة اللغات منْ خلفياتٍ ثقافيةٍ مُختَلِفةٍ وبالخبرة المكمّلة في عِلمِ النفْس "السائدِ" (يتضمن ذلك معرفة التراكيب ومقياسِها)، علم النفس القياسي، تقنيات تراكيب الاختبار، بالإضافة إلى الثقافة النفسية (المحلية) وعلم النفس المتعلق بالتقاطع الثقافي واللغويات. إن فائدة هذه الطريقة هي أنه تضمن الحد الأقصى من عدم

مـركـزية اللغـة والـثقـافـة فـي تعريف التركيب والاختبـار الذي صـمم لقيـاسـه. إن الخـواص المعينـة للغـة محـددة (ومـثـال، تعـابـيـر محليـة) أو ثقـافـة مـا (مـثـال، مـعـاييـر اجتماعية) يُمكنُ أنْ تُكتَشفَ وتُزالَ أثناء مراحل مبكّرةٍ مِنْ تطويرِ الاختبار؛ لذا، فإن طريقـة المعـالجـة المتـزامنة ذات دور فعّال فـي تطوير اخـتـبـارات صـادقـة مـتـعـددة الثقافة/ متعددة اللغة، بخصائص جيدة على حد سواء لكُلّ مِن الثقافاتِ المرجعية.

على أي حـال، هذا يَتـضمّنُ أكثـر مِنْ مـجـرد إنتـاج عـدة صيـغ لـغـوية لأداة مـا، وتجْميع البيـاناتَ في المجموعـاتِ الثقـافيةِ المُختـارةِ بشكل عشوائي ("دِراسـات رحلةِ صيد").

في طريقة تطويرِ الاختبـار المتعاقبةِ، بالممارسـة الأكثـر شيوعـاً، يُطوّرُ الاختبـار ويُصدّقُ بمطور، أو عدّة مطوّري اختبـار من خلفيـة معينة متعددة الثقـافـة/ متعددة اللغـة، وذلك الاستخدام ضمن هذا السيـاق متعدد الثقـافة/ متعدد اللغـة ("مصدر" لغة/ ثقافة). في أغلب الأحيان، بعد أن يصبح اختبـاراً شائعـاً بعدة سنوات، عندها فقط يتم تكييفه من قبل مطوري الاختبار الأصليين.، أو من قبل لجنة عمل جديدة للاستخدام في الثقـافـات الأخرى و/ أو إن علم النفس المختص بالتقـاطع الثقـافي يمكن أن يُؤدّي – دون قصد – إلى انحيـاز لتحيزات عرقية أو ثقافية تُحدّدُ التطويرَ اللاحقَ مِنْ النسخِ ذات الجودة المعادلة، في اللغات الهدف الجديدة و/ أو الثقـافات.

هذه المشكلة ستتفاقم بشكل أكبر إذا اشتركت لجنةُ عمل في تكيّفِ الاختبار لا تملك نطاقاً كامـلاً من الخبرة كما وصفت الطريقة الآنية. على الرغم من هذا، بمثل هذا العدد من الاختـبـارات الممتازة الأحادية الثقـافـة/ الأحـادية اللغـة، فـإن طريقـة المعالجة التي تعتمد على التعاقب، ستَستخدمُ كثيراً في المستقبلِ المنظورِ.

المخاطر والعلاج
في تطبيقات الاختبارات
المتعددة الثقافة/ المتعددة اللغة

ستُصادفُ الباحثين المختصين في تطويرِ الاختبارات المتعددة الثقافات/ المتعددة اللغات مشكلات ومخاطر غائبة في تطبيقات الاختبارات المتعددة الثقافات/ المتعددة اللغات.

باستعمال عدد مِنْ الأمثلة مِنْ موجوداتِ التقاريرِ الذاتيةِ وفحوصاتِ الأهلية غير الشفهية، نُصوّرُ تشكيلة واسعة لهذه المخاطرِ، وتحدد اختبارات دلائل التكييف ITC التي قَدْ تطبق عليهم، ونَقترحُ حلولاً محتملةَ. الأمثلة الأخرى للمخاطرِ والعلاجِ المحتملِ مقدمة في غرينفيلد (1997)، فان دي فيفر ولونغ، (1997b)، و فان دي فيفر وتانزر (1997).

مخاطر سببها المواد الوحيدة:

على خلاف التركيبِ وتحيّز الأداة/ إدارة، "وظيفة مادة تفاضلية" (سيرسي وآلالوف،2003) أو "تحيّز مادةِ" سببت بواسطة التشويهاتِ في مستوى المادةَ، فإن مادةً متحيزة لَها معان نفسية مختلفة عبر الثقافاتِ.

بالرغم من أن تحيّز المادةِ يُمْكِنُ أنْ يُنتَجَ من قبل بالعديد مِنْ المصادرِ، إلا أن السبب الأكثر تكراراً هو (أ) المقاييس الإنسانية مثل الترجمة الركيكة، أو الغموض في محتوى المادة الأصلية. أو من قبل (ب) "خواص ثقافية " أصيلة، مثل عدم الاطلاع/ تناسب محتوى المادةَ في بَعْض الثقافاتِ (توجيه 6). كما هو مُصَوَّرٌ في المثال التالي، إنه من الصعب في أغلب الأحيان في المواقع المتعددة اللغاتِ والمتعدّدة الثقافاتِ تَقْرير أي من هذين الاحتمالين يُمكنُ أنْ يبعد التفسيرات المختلفةَ و/ أو ردود الأفعال على الموضوعات.

محتوى المادةِ الغامض (المثال 1). في بيان حالة التعبير لنبرة الغضب، تعبيرِ غضبِ الميزةِ الرسميةِ (STAXI، سبيلبرغر 1988)، تم تمييز ثلاثة مِنْ أساليبِ تعبيرِ

الغضبِ. غضب موجه للخارج، وهو تعبيرُ الغضبِ نحو الناسِ أو الأشياء الأخرى. غضب داخلي، وهو إخمادُ المشاعرِ الغاضبةِ، وغضب مسيطر عليه، وهو المحاولة للسيطرة على تعبير الغضب. عموماً، تراكيب العوامل الثلاثة المُفتَرَضةِ لتعبيرِ الغضبِ أُكّدَت في عدّة نماذج أمريكية (سبيلبرغر، 1988) نموذج لصينيين سنغافوريينِ (تانزر، سيم، سبيلبرغر، 1996) وكذلك في التكييفِ إلى الألمانية (سكوينكميزجر، هوداب سبيلبرغر، 1999)، الإيطالية (سبيلبرغر وكومونييان، 1992) والنرويجية (هاسيث، 1996) على أية حال، في بَعْض الدراساتِ (مثال، سبيلبرغر وكومونييان، 1992، تانزر و سيم و سبيلبرغر، 1996)، إن المادة "أنا ناقد للآخرين تماماً بشكل سري " قد انتقلت من الغضبِ الداخلي إلى الغْضب الخارجي.

تعريف المشكلةِ. من المُهمِ مُلاحَظَة (التوجيه 13) بأنَّ هذه المادة يُمْكِنُ أنْ تُترجمَ إما كـ" إخفاء ضغائن وعدم التحدث عنها للناس الآخرين" الذي سيكون عندها تعبيراً عن غضب داخلي. أو يمكن ترجمتها كـ " التحدث بشكل سلبي بغياب شخص ما" وهذا قد يحمل تعبير غضب خارجي خفي، بدلاً من ذلك. في هيئة لدراسة التقاطع الثقافي، تم إعطاء طلبة كليات ناطقين بالألمانيةَ والإيطاليةَ بيان (STAXI بيان حالة التعبير لنبرة الغضب) بصيغة لغتهم الأصلية.

في إعادة الاختبار، سئلوا كيف سيُترجمونَ المادةَ. بالرغم من أن حوالي نِصْفِ الطلابِ الناطقين بالألمانيةِ 53 ٪، (n=347) اختاروا البديل الأولَ، ثُلث فقط مِنْ طلبةِ الكليات الناطقين بالإيطاليةِ 64 ٪، (n=241) اختاروا هذا البديلِ. هكذا، بإعطاء غموض لمحتوى المادةَ، حتى الاختلافات البسيطة بين صيغ اللغةِ المختلفةِ يُمْكِنُ أنْ تُسبّبَ نقلات كبيرة في انتشار التفسيراتِ البديلةِ (توجيه 7).

بالرغم من أنَّ تعبير الغضب والتغلب عليه مشاعر عالميةً، لكنها على الرغم من هذا هي محكومة بالعواملِ الثقافيةِ (مسكيتا وفريجدا، 1992). على سبيل المثال، إن المفاهيم "الحاجة إلى الانسجام "و" تَعطي وجهاً" سائدة في المجتمعاتِ الصينيةِ

(ومثال على ذلك: -، بوند، 1990؛ شونغ وآل، 1996؛ غاو، 1998) تُمْنعُ المجابهات المفتوحة والمباشرةَ التي تُميّزُ الطرقَ الأمريكيةَ للتعبير عن الغضبِ الموجه للخارج.

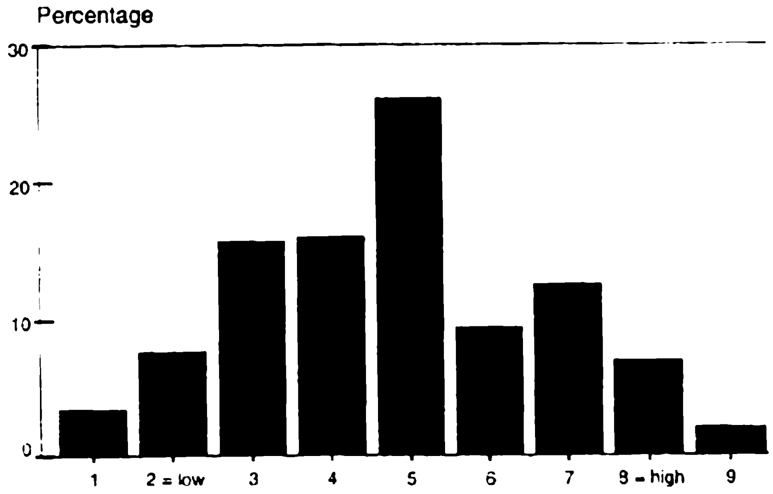
هكذا، نمط غير تجابهي غير مباشر للتعبير عن الغضب كما هو متضمن في التفسيرِ الثاني الذي سيتم تفضيلهُ. في هذه الحالةِ، إن النقلات المُلاحَظة تَعْكسُ اختلافات ثقافية أصيلةَ في طريقة إبداء تعبير الغضبِ الموجه للخارج (توجيه 3).

المواد المثبتة الجدلية (المثال 2). إن استبانه إجهادِ العمل JSS، سبيلبرغر وريهيسر، (1994) هو تقرير ذاتي صمّم لقيَاس تأثيرِ الإجهادِ المتعلّق بالعمل. ويتألف من 30 بياناً الذي يصفف أحداثَ متعلقة بالعمل وصفت بالمرهقة مرهقة من قبل مستخدمين في وظائفِ مُختَلفةِ (اجتهادات Job Stressors) وكُلّ ضغط عمل تم تقديره مرتين: تَقدير متعلق التكرار الذي حدث فيه خلال الستة أشهر الأخيرة (تقدير التكرار). أما الثاني فهو كمية أو درجة الإجهادِ الذي أُثير بسبب حدث واحد من الأمور المسببة للضغط (معدلات "الشدة" Severity ratings). ومن أجل تقديرات الشدة، تم توجيه الموضوعات لمُقَارَنَة كميةِ الإجهادِ المترافقة مع كل مسبب للضغط، بكميةِ الإجهادِ المثار من قبل مسبب للضغط " مهمة الواجبات المزعجة"، تم اختيارها كمرساة القيمة المسبقة التخصيص للشدّةِ المتوسّطةِ.

تعريف المشكلة. أثناء تكييف استبانة إجهاد العمل JSS، ، تم دراسة الترجمة الألمانية للمثبت بتحويل هذا الإجهاد إلى مادة حرة التقدير (التوجيهات 8، 9، 13). على الرغم مِنْ معدل التقديرات المتوسط، تلقت هذه المادة تقديرات جدلية بثبات في عدد مِنْ الدراساتِ بغض النظر عن صيغ الرَدّ والتعليمات المستعملة (هودابو تانزر، ماير، بيستيمر، وكورونكا، في الصحافةِ). علاوة على ذلك، تم إيجاد ثبات عال متعلق بالاختبار وإعادة الاختبار لهذا الإجهاد في دراسة لائحة ($r_{tt} = .68; N = 156$) أشارت بأنّ الفرقَ الواسعَ للتقديراتِ (انظر الشكل 10) لا يمكن أن يعزى ببساطة إلى التأثيرات العشوائية. في الدراسةِ الأخرى التي استعملت هذه المادة كمرساة مع قيمة مسبقة

التخصيص لشدة متوسطة، لوحظ أن عدداً من الموضوعات لَمْ تمتثل إلى الأمر. عندها ألغوا تَقدير الشدّة المتوسط المسبق التخصيص، وصنفت كفئة تقدير مختلفة بدلاً من ذلك. باختصار، إن النتائج التي تشير إلى أنّ هذه المادة لَمْ تُستدعَ كمية الإجهاد نفسها لكُلّ المستجيبين الناطقين بالألمانية، يعد طلباً ضرورياً قبل أن يمكن تأسيسه كمادة مثبتة صادقة.

الحل المقترح. اجمع "معلومات عرضية" (مثال، تفسير المتحنين لمحتوى المادة في المثال الأول، وطرق غير قياسية مِنْ إدارة الاختبار في المثال الثاني) لاختبار صدق الاختبار المتعدد الثقافة/ المتعدد اللغة كمادة مثبتة (التوجيه 14).



الشكل رقم 10-1

توزيع معدلات الشدة لإجهادات العمل في استبانة إجهاد العمل «تحديد المهمات غير المتفق عليها» على مقياس من ٩ نقاط.

مخاطر سببها تصاميم الاختبار غير المتوافقة ثقافياً

هناك رأي واسع الانتشار بين العلماء النفسانيين السائدين، وهو أن المشكلات والمخاطر التي تمت مصادفتها أثناء تطبيقات الاختبار المتعدد اللغات، سببها بشكل رئيس استعمال المادّة الشفوية (مثال، في بيانات التقارير الذاتية أو اختبارات القدرة الشفوية مثل مقياس ذكاء فكسلر للأطفال/ مقياس تكييف ذكاء فكسلر للبالغين)

بشكل مكثف. ولذلك استخدام اختبارات "لا شفهية" مثل اختبارات جداول رافن المتعلقة بالمقاييس التقدمية (SPM) في تقييم القدرة، أو الاختبارات "التصويرية" في تقييم الشخصية، سيخفض بشكل كبير مشكلة تطوير أداة صادقة للتقييمات المتعددة الثقافات/ المتعددة اللغات. تُبيّنُ الأمثلةُ التالية بأنّ هذه لَيستْ بالضرورة القضيةَ.

التوافق الثقافي للمادّةِ التصويريةِ (المثال3). يستخدم عدد مِنْ الاختبارات في الشخصيةِ و الصحة العقليةِ، محفّزاتِ تتألف بشكل خاص مِنْ مادّةِ تصويريةِ. الأمثلة هي اختبار إبطال الصورة لروزينزويغ (مثال، روزينزويغ، 1977)، علامات ما قبل المدرسة للتقرير الذاتي (مارتيني، سترايهورن وبويغ-انتيك، 1990) و والتقدير التصويري لردات فعل الاختبار توبيانا، (1994) باعتبار أن كل المولد تشمل عناصر محددة لثقافة معينة (مثال، الأسلوب الغربي لارتداء الملابس، تصفيفات الشعر أو الملابس الثنائية الجنس). مدد الوقت أو مجموعة الأخلاقيات (مثال، وجوه بيضاء). إن تطبيقاً صادقا لتقاطع الثقافات، يَتطلّبُ إعادة رسم كاملة لكُلّ المادّة التصويرية (التعليمات 1، 3، 6). بناء على ذلك، يجب أن يتم إنشاء اختبار جديد كلياً، ويَجِبُ أنْ يُصدّقَ لكُلّ ثقافة جديدة.

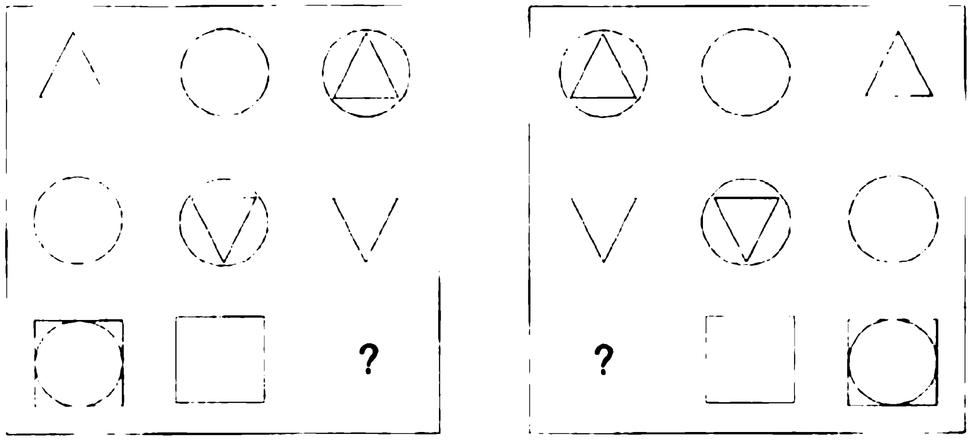
كتابة التعليمات (مثال 4). في دراسة للتقاطع الثقافي (بيسوانجر، 1975، انظر أيضاً إلى فيشر و فورمان، 1982) واختبار القوالب الفانيسي VMT (Viennese Matrices test)، فورمان وبيسوانجر، 1979 قدمت إلى طلابِ المدارس العليا العربِ والنايجيريينِ والتوجوليينِ.. وقد تم مقارنة إجاباتهم مع آخرين من نموذج المعايرة الأسترالي. إن VMT هو اختبار للمنطق الاستقرائي يستعمل أسئلة جداول مماثلة لـ SPM الخاص برافن. وقد أظهر الاكتشاف الأكبر المفاجئ أن تحديد وتطبيق القاعدة من اليسار إلى اليمين كمقارنة للأعلى وللأسفل، كان أكثر صعوبة للطلاب العرب الأفارقة مِنْ الطلابِ النمساويينِ.

تعريف مشكلة. إن المزاوجة على نحو غير ملائم بين اتجاه كتابة في الثقافة العربية (وبمعنى آخر: من اليمين إلى اليسار في العربية، يقابلها من اليسار إلى اليمين في اللغة اللاتينية) والطريقة الغربية لتصميم أسئلة الجداول بالعنصر المفقود في الزاوية اليمين السفلية (انظر إلى الجانب اليساري من الشكل. 10-2) يعد تفسيراً معقولاً. لتوضيح النتائج الإدراكية التي يمكن أن يكون سببها تصميم الأسئلة غير المتوافقة ثقافياً، تخيل ارتباك عرب يقرؤون إعلاناً مؤلفاً من ثلاثة صور لآلة غسيل. عند ترتيب قراءتهم من اليمين إلى اليسار نجد امرأة (أ) تبدو سعيدة عند كومة من الغسيل النظيف. (ب) تضع الغسيل في الآلة وتديرها، ثم (د) بعد ذلك تفزع من أكوام الغسيل القذر.

الحل المُقترح. إن استعمال تصميم منسجم مع الثقافة لأسئلة الجداول، قد يتغلب على هذه المشكلة (التوجيهات 5، 14، 15). كما هو موضح في جهة اليمين للشكل 10-2، صيغة الاختبار "معكوسة" حيث تم ترتيب عناصر الجدول بطريقة يكون فيها العنصر المفقود عند أسفل زاوية اليسار، قد يزيل هذا النوع من التحيز للممتحنين العرب.

معنى المريكات. (المثال 5). مثال آخر لتصميم المادة غير المتوافقة ثقافياً تم تقديمه من قبل غرينفيلد (1997). إن معنى الإرباكات في الاختبارات المتعددة الخيارات يمكن أن تُفسر بواسطة موضوعات من عدة ثقافات بشكل مختلف (التوجيهات 5، 14):

في الصيغة المتعددة الخيارات، أعطي المجيب مجموعة من الخيارات. كلها باستثناء واحد، وُظفت لتكون معلومات عديمة الجدوى. المشكلة هي أن المشاركين من عدة ثقافات سوف يفترضون أن واضع الاختبار يقدم مجموعة من المعلومات تتعلق بهدف حل المشكلة...... إن واضع الاختبار (ومطوره) يفترضان أن هدف الإجابات البديلة هي إزالة الاحتمالات غير الصحيحة، أثناء اختيار البديل الصحيح. يفترض الممتحن أن هدف الإجابات البديلة هي إنشاء حل للمشكلة. (ص: 1120-1121).



معكوس   نموذجي

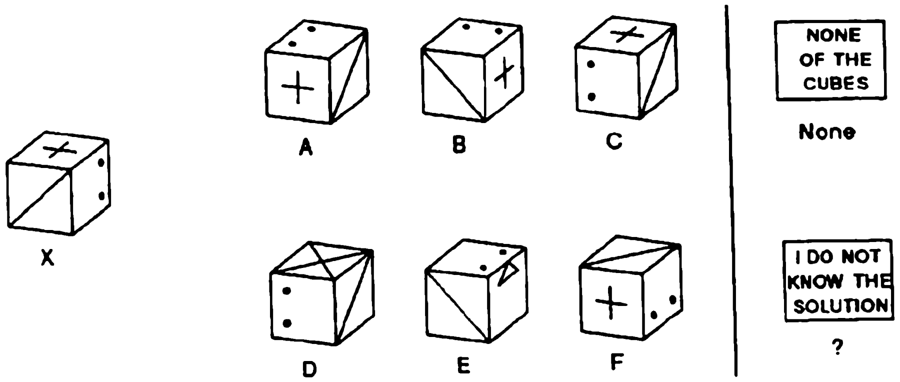
الشكل 2.10

بند اختبار معياري (بدون أسئلة الخيارات المتعددة التي تصرف الانتباه) في حاليتها النموذجية والمعكوسة

الحل المُقترح. كما أشار غرينفيلد (1979، تحمل المشكلة الحل بذاتها. لا تستعمل صيغ اختبارات متعددة في التطبيقات المتعددة الثقافات/ المتعددة اللغات، إلا إذا كان كل الممتحنين مطلعين بشكل كامل على مفهوم المربكات. بدلاً من ذلك، دع الممتحنين يستنتجون (مثال: استدرج) بدلاً من اختيار الحل.

تأثيرات الممارسة الضمنية للاختبار. (المثال 6). معتمدين على معارفهم وخبراتهم السابقة بالموضوعات المشابهة لأسئلة الاختبار (التوجيه 6)، سيستفيد الممتحنون بشكل مختلف من تأثيرات الممارسة الضمنية للاختبار. خاصة في بداية الاختبار (توجيه 14). أُخذت الصورة الإيضاحية التالية من دراسة متعلقة بتقاطع الثقافة لاختبار مقارنة المكعب الثلاثي الأبعاد 3DC، غيللر، (1990) وهو اختبار مقارنة مكعب يقيس القدرة المكانية. نجد مثالاً لسؤال منه في الشكل 3.10 هذه الدراسة المتقاطعة الثقافة (بروير، 1996، تانزر وغيتلر وإيليس، 1995، تانزر وغيتلر وسيم، 1994) أشارت إلى أن الطلاب القادمين من البلاد التي ليس لديها تعليم رسمي في الهندسة الوصفية، مقارنة بنظرائهم النمساويين، حصلوا أكثر من العمل على أسئلة الاختبار القليلة الأولى.

الحل المقترح. استعمل عدداً كافياً من الأسئلة الإحمائية الخفية، أي أسئلة مقدمة كالأسئلة الحقيقية في بداية الاختبار ولكن ليس لاستخدامها من أجل الدرجات. في صيغة المصدر الأحادي الثقافة لـ 3DC غيللر، (1990)، تم استعمال الاختبار الأول كسؤال إحمائي مخفي، بينما استعملت أسئلة الاختبار الـ 17 المتبقية للدرجات. ولكن دراسات التقاطع الثقافي المذكورة سابقاً، قد أشارت الى أنه يلزم المزيد من الأسئلة الإحمائية الخفية، لتأمين مقارنات درجات صادقة في السياقات المتعددة الثقافات (توجيهات 20-12).



الشكل 10-3

نموذج بند استخدم في إرشادات 3DC الموجودة في الملحق

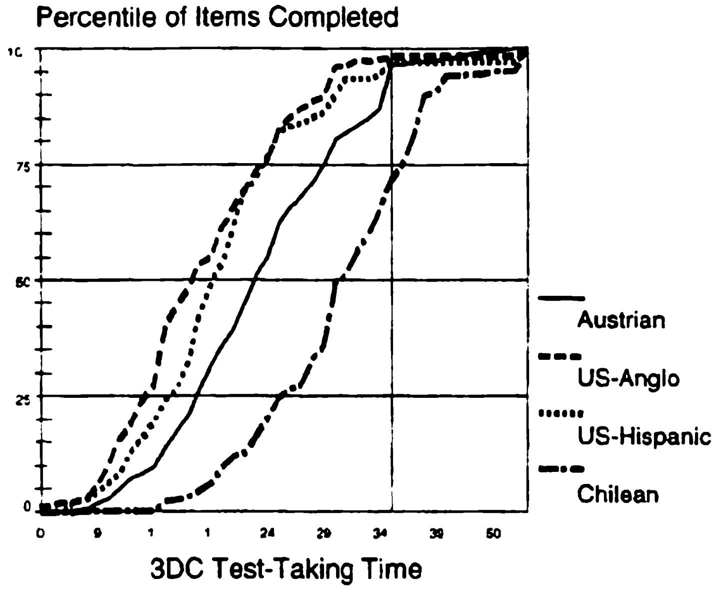
لاحظ أنه لا يمكن أن تحل هذه المشكلة بمجرد تقديم المزيد من أسئلة الممارسة مع حلولها المعطاة؛ وذلك لأن معظم الممتحنين يعالجون أسئلة الممارسة بشكل مختلف عن أسئلة اختبار حقيقية. يمكن تحديد العدد الضروري من الأسئلة الإحمائية الخفية، عبر إعطاء الممتحنين بشكل عشوائي صيغة اختبار نموذجية مقابل صيغة اختبار بترتيب أسئلة معكوس، ثم مقارنة صعوبات السؤال بين صيغتي الاختبار (توجيهات 8، 9، 22).

المخاطر المسببة بواسطة تعليمات الاختبار العرقية وإجراءات الإدارة

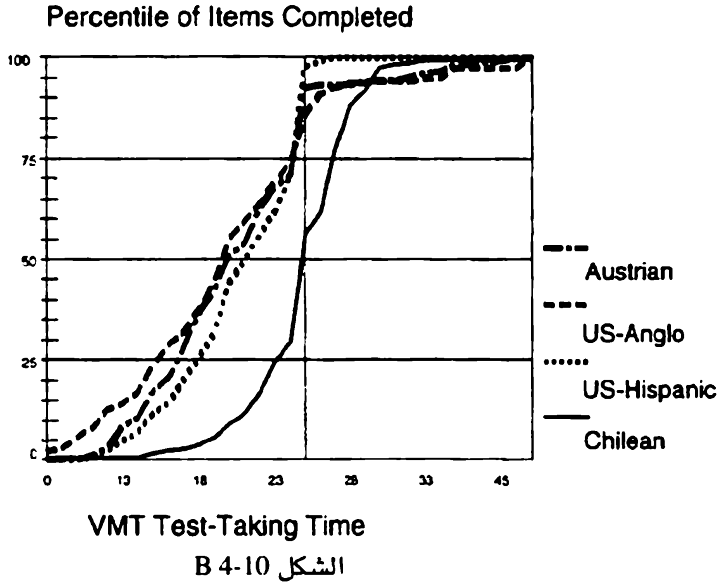
مصدر آخر للتحيز العرقي هو اختبارات الذكاء/ الأهلية غير الشفهية التي قد تكون مخفية في تعليمات الاختبار وإجراءات الإدارة. على سيبل المثال، إن نقاط أفضلية الممتحنين على منحنى الخط البياني لتبديل الدقة والسرعة، قد يتأثر بقيم ثقافية معينة متعلقة بالوقت والدقة.

فرض حدود الوقت في اختبارات القدرة (المثال 7).

على الرغم من أن VMTو 3DC قد صممت كاختبارات استطاعة (أي الكفاءة الذاتية)، فإن محرري الاختبارات قد أوصوا بحدود زمنية متسامحة لإدارة اختبار اقتصادي. في دراسة التقاطع الثقافي مع هذين الاختبارين (برויר، 1996، تانزر وإيليس وغيلر وبروير، 1996، تانزر وآل)، فإن حد الوقت الموصى به والمقدر بـ25 دقيقة لـ VMT و35 دقيقة لـ 3DC، قد تم زيادته لإعطاء كل الممتحنين وقت اختبار متسعاً (توجيه 13). كان المشاركون طلاب جامعات أو كليات من النمسا (N=244)، تشيلي (N=173)، والولايات المتحدة مع أنجلو- أمريكي واحد (N=196)، ونموذج إسباني واحد (N=144) كما يمكن رؤية ذلك في الشكل 4.10، فإن نسبة كبيرة من الممتحنين التشيليين قد تجاوزا كلاً من حدي الوقت الموصى بهما.



الشكل A4-10



الشكل B 4-10

النسبة المئوية للأشخاص الذين أكملوا اختبار 3DC (الشكل A4-10) واختبار UMT (الشكل B4-10) مرسومة بالمقابل لزمن الاختبار (بالدقائق للنمساويين (N=244) والتشيليين (N=173) والإنجلو أمريكان في أمريكا (N=196) والهسبانيك في أمريكا (N=144).

بالتالي، فإن فروقات التقاطع الثقافي التي تم تحصيلها تحت تعليمات نموذجية ستكون مهمة. وبذلك، تشير بشكل غير صحيح إلى أن المجموعات تختلف في مهارات المنطق الاستقرائية (التوجيهات 14-22). باعتبار أن حدود الوقت في اختبارات القدرة يمكن أن تضر بمجموعات ثقافية معينة، فنحن نوصي بشدة تسجيل الوقت المستغرق في الاختبار وذلك للاستخدام في "المعلومات الإضافية" (توجيه 20) في تطبيقات اختبارات التقاطع الثقافي لاختبارات القدرة.

قابلية ترجمة تعليمات الاختبار (المثال 8). إن التعليمات في الاختبارات الأحادية الثقافة/ الأحادية اللغة، غالباً ما تنسجم بشكل جيد مع التطبيقات المقصودة الأحادية الثقافة/ الأحادية اللغة. وهكذا، فهي تميل إلى استغلال الخاصية اللغوية (التبحر في اللغة). وقد قام بريسلين (1986) بصيغة عدد من التوجيهات لتحسين قابلية الترجمة للمواد الشفهية التي يمكن أن تطبق أيضاً على صيغ تعليمات الاختبار (التوجيه 4).

توثيق ما بعد التعليمات (المثال 9). عادة ما تُطور تعليمات الاختبار عبر سلسلة من النسخ و" المعرفة-الكيفية" التي تم إحرازها في أثناء عملية (مثال، جزء معين من المعلومات، عناصر مهمة للتعليمات، خلفية تعليمية مطلوبة للفهم، إلخ) التطبيق في النسخة النهائية. ولكن معرفة الكيفية هذه نادراً ما توثق. وهكذا، عادة ما تكون غير متوفرة للجنة العمل المعنية في التكييفات التسلسلية للاختبار *هذه*، مع النسخة النهائية للتعليمات (التوجبه 17). وهذا يمكن أن يطبق كتوثيق معلومات مخزنة في حاسوب (مثال، كونكلين، 1987)، كما هو موضح في فهرس التقييمات الإنكليزية لـ DC 3 .

التكييف المتعدد الثقافات/ المتعدد اللغات لاختبار التركيز (المثال 10). إن كلاً من الأمثلة السابقة قد ركز على خطر واحد فقط. لإلقاء الضوء على عدد من المشكلات، قد يواجه المرء، حتى بالاختبارات البسيطة نسبياً، المثال التالي الذي يقدم صورة عن تكييف الاختبار المتعدد الثقافات/ اللغات.

النسخة الأصلية الألمانية. إن d2 بريكن كامب وزيلمر،1998 هو اختبار Bour-don الذي يأخذ شكل رسالة مختزلة. وهو مصمم لقياس التركيز القصير الأمد. تم تطويره في الأصل في ألمانيا (بريكن كامب، 1962) وما يزال يستخدم بشكل متكرر في البلاد الناطقة باللغة الألمانية. يتوجب على الممتحنين أن يقوموا بمسح 14 صف hW من d's و p's مع القواطع فوق و/ أو تحت الإشارات المثبتة بأسرع ما يمكن دون ارتكاب أخطاء. إن أنواع الإشارات الثلاثة هي (أ) d's مع قاطعتين فوق، (ب) d's مع قاطعتين تحت، و(د) مع قاطعة فوق وقاطعة فوق. تتألف مادة الاختبار الأصلية من ورقة اختبار مصورة مزدوجة الجهة، مع تفصيلات شخصية، أحرف الإشارات، أسطر كمثال، وجدول العلامات اليدوية على الصفحة الأمامية (الصفحة اليمنى) و4 سطر للاختبار مع مؤشرات العلامات على ظهر الصفحة (صفحة اليسرى). تتم إدارة d2 ضمن شروط السرعة (مثال، مُؤقت) بحوالي 4 دقائق لوقت الاختبار. وتعطى كل التعليمات بشكل شفهي.

تطويرات متعاقبة لنسخ اختبار لا مركزي:

بدأ فريق من علماء النفس المتعددي الثقافة/ المتعددي اللغة (تانزر، إيليس وآل، 1997) بتطوير صيغ اختبار لا مركزية ثقافياً ولغوياً. ويجب أن تكون صادقة تحت الشروط التالية المحددة مسبقاً:

(أ) يجب أن يكون لدى كل الممتحنين 8 سنوات على الأقل من التعليم الرسمي، (ب) يمكن أن لا يكون لدى الأغلبية منهم خبرة مسبقة باختبارات Bourdon المختزلة، (د) يمكن أن يجري الاختبار ضمن مجموعة صغيرة حتى المجموعات المتوسطة الحجم، (ج) لا يجب أن تعطى التعليمات باللغة الأصلية للممتحنين، (ح) يمكن أن يكون واضعو الاختبار: (1) علماء نفس أجانب مدربين، يملكون خبرة واسعة في التقييمات النفسية وفي إدارة d2، ولكن بلهجة أجنبية قوية وخبرة قليلة نسبياً بالظروف المحلية. (2) علماء نفس محليين مدربين، مع معرفة واسعة

بالظروف المحلية وخبرة في التقييمات النفسية بشكل عام، ولكن ليس بإدارة d2 بشكل خاص. أو (3) طلاب في مرحلة التخرج مع خبرة متنوعة في الاختبارات النفسية.

استناداً إلى الإجراءات الحكمية (توجيه) والبرهان التجريبي في سلسلة من دراسات التقاطع الثقافية، مع نماذج من 150-250 لطلاب كليات أو جامعات، تم تطوير الصيغ اللامركزية للغة الصينية، اللغة الكرواتية، اللغة الإنكليزية (المستخدمة في الولايات المتحدة ومن قبل النماذج الأنجلو أمريكية والإسبانية)، اللغة الفرنسية (المستخدمة من قبل النماذج النمساوية) ولهجات إسبانية متعددة. بافتراض السهولة في d2، قد يبدو أن الترجمة الصحيحة لتعليمات الاختبار ستكون كافية. ولكن، كما هو موضح لاحقاً فإن عدداً من التغيرات في تصميم مادة الاختبار، تعليمات الاختبار وإجراءات الإدارة، كانت ضرورية من أجل ضمان تكافؤ التقاطع الثقافي واللغوي.

تصميم ورقة الاختبار (المثال A10) تم تبسيط النموذج الطباعي للاختبار الأصلي عبر تحريك جدول الدرجات اليدوية وكل حقول البيانات الحياتية باستثناء الرمز الشخصي من تعليمات الصفحة (الصفحة اليمنى) ومؤشرات الدرجات من صفحة الاختبار (الصفحة اليسرى). من أجل التأكيد على أن كل الممتحنين سوف يفهمون التعليمات تحت كل الشروط المذكورة مسبقاً، تم تقديم الجزء المركزي من التعليمات بصيغة مكتوبة على صفحة التعليمات. هذا يمكننا أيضاً من تحريك التفسيرات " ك: كما في كلمة كلب" و "خ: كما في كلمة خنزير" من التعليمات "الشفهية " المقدمة في النسخة الإنكليزية (بريكن كامب وزيلمر، 1998) وهو شيء مربك أكثر منه مساعد في الثقافات التي لا تستخدم حروف الأبجدية اللاتينية (مثال، الصين). في النهاية تم إبراز حروف الأهداف الثلاثة في مربع عُلِم مرتين على سطر النموذج.

الإدارة (المثال، B10) لضمان إدارة ثابتة تحت شروط الحقل التي تم ذكرها سابقاً، جرى تغيير الوقت المجزأ من 20 دقيقة لكل من السطور الـ 14 إلى دقيقتين لكل من مجموعتي (جزءA وجزء B) من السطور السبعة. بالإضافة إلى ذلك، تم وضع قائمة مفصلة من التوجيهات لإدارة الصيغة اللامركزية في السياقات المتعددة الثقافة (التوجيهات 8، 17).

التعليمات (المثال، C10) باعتبار أن عدداً من المختبرين استعملوا طرقاً مختلفة لتعليم الأهداف (مثال، الدرجات الصغيرة. التقاطعات، الخربشات والدوائر)، تم تأكيد وتبرير ("لتوفير الوقت") الطريقة المطلوبة " للتعليم بشحطة واحدة"، ووضحت مرتين في سطر النموذج. بطريقة مماثلة، تم معايرة الطرق المختلفة لتصحيح التعليم الخاطئ (مثال، شطب، خربشة، محي بالمحاة) إلى "الشطب" (X-out في النماذج الأمريكية)، ووضحت على السبورة أو على الزجاج المضاء.

وُجه أيضاً اهتمام خاص إلى ترجمة مفاتيح العبارات. في النسخة الإنكليزية، تمت الإشارة إلى الأنواع الثلاثة من الإشارات المحتملة (أي: كل الـ d's مع مجمل التقاطعين) مثل " أمثلة" (بريكن كامب وزيلمر، 1998) ولكن لأن تعبير: أمثلة لا يحمل معنى مجموعة مستنفذة، تم تغييرها إلى "هدف". علاوة على ذلك، فإن الصيغة الغامضة لـ:شطب" استخدمت بشكل متكرر (مثال، "لا يفترض شطب الحروف الأخرى" " الحرف خ كما في خنزير" يجب أن لا تشطب أبداً) في التعليمات الشفهية. إن الطريقة المطلوبة ("للشطب...... بإنشاء خط واحد عبر الحرف") تم تحديدها فقط مرة واحدة. باعتبار أن التعبير "شطب" يمكن أن يفسر إما كطرقة معينة للتعليم (أي، الاستبعاد " X-out") أو كمفهوم أكثر عمومية للحذف/ الإلغاء، يتضمن أنواعاً أخرى من التعليم ("حذف"، "محو")، فقد تم استبداله بعبارة أكثر دقة (التعليم بشحطة واحدة). أخيراً، إن عبارة "توقف" تابع مع B، التي أعطيت بعد أول دقيقتين قد جعلت بعض الممتحنين يتوقفون بشكل حرفي عن الحل وهكذا، تم استبدالها بـ " الوقت" تابع مع B"

حدود تكييفات الاختبار المتتابعة: بينة للتطوير المتزامن

في كل الأمثلة المقدمة، لم يتمكن مترجمون محترفون (تطبيقات الاختبار، فان دي فيفر ولونغ. 1997a و1997 b) من تقديم أدوات صادقة متعددة الثقافات/ متعددة اللغات. حتى في حالة الاختبار البسيط نسبياً مثل d2 (مثال 10)، كان عدد من التعديلات الجوهرية ضرورياً " لتكييف الاختبار. قد يتطلب تكييف الاختبارات تعديل سؤال واحد (مثال 1)، وصيغ الإجابة (الأمثلة 2 و5)، وتصميم الأسئلة (مثال 4) وورقة الاختبار (مثال 10a)، وقواعد الدرجات (مثال 6) وإجراءات الإدارة (الأمثلة 7 و 10b)، وتعليمات الاختبار (الأمثلة 8، 9، 10c) وفي حالات أخرى (مثال 3)، قد يكون من الضروري تكييف الأداة إلى حد وضع أداة جديدة تماماً "تركيب اختبار" بشكل خاص.

في الواقع، على الرغم من الجهود الكبيرة لإنجاز صيغة ثقافية ولغوية لامركزية لـ d2، تبقى المشكلتان الجادتان التاليتان اللتان يمكن حلهما بإكمال إعادة تصميم الاختبار (أي تركيب الاختبار).

حروف أهداف لاتينية (المثال 11). إن استخدام الحروف اللاتينية (d's و p's) في d2 يمكن أن يضر المختبرين الذين لا يستخدمون حروف الأبجدية اللاتينية في حياتهم اليومية. من جهة أخرى، بإمكان المختبرين المدُربين على حفظ وتمييز الأحرف المعقدة مثل أحرف اللغة الصينية الشمالية، أن يقوموا بغربلة السطور بوقت واحد للمراحل الثلاثة بدلاً من مسح d's أولاً ثم مراجعة عدد التقاطعات. لاحظ أيضاً أن اختبارات إلغاء الحرف هي أمثلة نموذجية لتراكيب الاختبارات العرفية لأنه ليست هناك حاجة فعلية للاعتماد على هذا النوع من المنبهات (غيتلر وتانزر، 1990/1998، موسبراغر أولسشلاغيل، 1996).

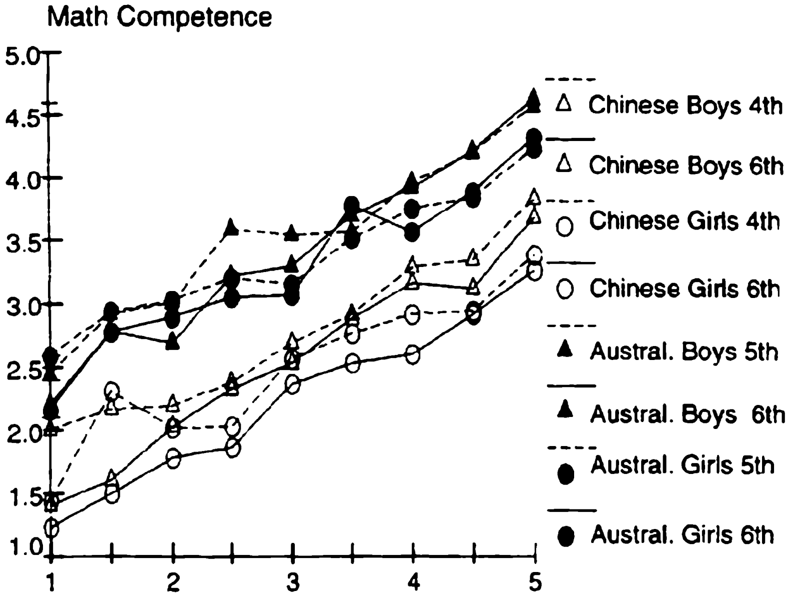
وجهة العمل (المثال 12). لم يمتثل عدد من الممتحنين من النماذج الغربية للتوجيه " اعمل على حل السطور من اليسار إلى اليمين" بل حلوا بشكل متعرج بدلاً

من ذلك. لم تلغ هذه الظاهرة أيضاً بالتوجيه الإضافي " قم بالحل كما لو كنت تقرأ كتاباً" (استخدم في النماذج الغربية). وهكذا، لم يكن من المحتمل أن هذا نتج عن سوء فهم التعليمات. بل كان السبب، على الأغلب، الممتحنين الذين يستعملون اليد اليمنى. وهؤلاء، بخلاف الذين يستعملون اليد اليسرى، تمت إعاقتهم بهذا الاتجاه من الحل؛ لأن يدهم غطت بقية الأحرف. تفاقمت هذه المشكلة في سياق متعدد الثقافة، لأن وجهة الكتابة الموصى بها قد أضرت بالممتحنين العرب (انظر أيضاً إلى المثال 4).

إن الخطر الموضح في الأمثلة السابقة له أثر فقط على صدق التقاطع الثقافي للقياس (أي الاختبار)، ولكن ليس على صدق التقاطع الثقافي للتركيب الأساسي. يوضح المثال التالي أن "تحويل الاختبار" (غرينفيلد،1997) من ثقافة أحادية إلى سياق متعدد الثقافات، يمكن أن يؤثر حتى على تكوين المفاهيم عن التركيب الأساسي (التوجيه2).

العناصر الضمنية للمفهوم الذاتي الأكاديمي (المثال 13). تكشف أبحاث المفهوم الذاتي الأكاديمي باستمرار أن القراءة وتقارير المفاهيم الذاتية الرياضية شيئان غير مترابطين (أي مجال محدد). في معظم هذه الدراسات، تم قياس المفهوم الذاتي الأكاديمي بواسطة الاستبيان الوصفي للذات (SDQ-I، مارش، 1988) الذي استخدم معايير تتألف من أسئلة إدراكية وعاطفية/ محرضة لقياس المفهوم الذاتي الأكاديمي في مجال محدد. في مقارنة للتقاطع الثقافي لطلاب نمساويين وسينغافوريين وصينيين (تانزر، 1995، 1998، تانزر وسيم، 1991 وتانزر وسيم ومارش، 1992، 1997) برزت اختلافات تتعلق بالتقاطع الثقافي في النسب المتبناه لأسئلة الكفاءة/ سهولة المهمة، ولكن ليس في أسئلة الاهتمام والاندفاع (انظر إلى الشكل 10-5). باعتبار أن النسخة الإنكليزية الأصلية لـ SDQ-I استخدمت في كلا البلدين، "تقييم متعدد اللغات/متعدد الثقافات"، فإن مشكلات الترجمة ليست بالتأكيد سبب هذه الظاهرة (التوجيهات 16.7).

إن تحليلات عوامل التداخل الثقافية اللاحقة في أستراليا (مارش، كرافن وديوس، 1999) النمسا وإيطاليا وسنغافورة والولايات المتحدة (تانزر، 1998) قد دعمت أيضاً فصل العناصر الضمنية الإدراكية (أي الكفاءة، سهولة المهمة) والعاطفية/ المحرضة لقياس المفهوم الذاتي الأكاديمي في مجال محدد. وهكذا، إن دمج أبحاث التقاطعات اللغوية والتداخلات اللغوية (أي، علم النفس الفارقي Differential Psychology) قد ساعد على اكتساب فهم أعمق للمفهوم الذاتي الأكاديمي.



الشكل رقم 5.10

انحدار جدارة الرياضيات في الاهتمام بالرياضيات ضمن المجموعات الثمانية الفرعية (العمر - الجنس - الثقافة) [الصينيون السنغافوريون مقابل الإستراليين]

يوضح هذا المثال أيضاً الحدود لطرق معالجة تطوير الاختبارات المتعاقبة: إن نتائج جديدة للتقاطعات اللغوية والتداخلات اللغوية تتعلق بتشكيل مفاهيم حول التركيب ومقاييسه، يمكن أن يتم دمجها فقط بلغة الهدف (أي المترجم إليها) ولكن ليس إلى نسخة المصدر. مثلاً، دَمج تكييف SDQ-I الألماني بشكل كامل مفهومي العنصرين للمفهوم الذاتي الأكاديمي في مجال محدد، بينما تقدم النسخة الإنكليزية الأصلية معايير "سابقة" فقط لهذه الغاية. علاوة على ذلك، فإن الجزء الكبير

للنظرية والبحث الذي جرى بواسطة الأداة الأصلية (مثال، بحث SDQ-I للمفهوم الذاتي الأكاديمي في مجال محدد، انظر الى مارش 1998، مارش وكرافن، 1997) قد لا يحافظ بشكل كامل على تركيب إعادة تشكيل المفهوم.

تصميم الاختبارات للاستخدام في الثقافات واللغات المتعددة:

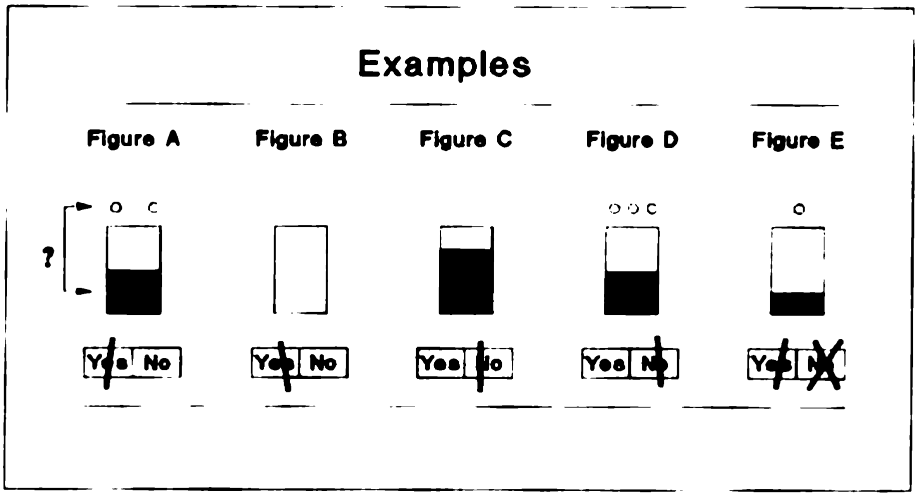
بافتراض قصور تكييف الاختبارات المتعاقبة، نحن نوصي بتطوير متزامن للاختبارات المتعددة اللغات والثقافات. بخلاف طريقة المعالجة المتعاقبة، يسمح تطوير الاختبار المتزامن بدمج النتائج من دراسات الثقافة المتقاطعة والثقافة المتداخلة؛ وذلك بإعادة صياغة إعادة تشكيل المفهوم للتركيب ومقياسه في مرحلة مبكرة نسبياً من تطور الاختبار. يمكن أيضاً لطريقة المعالجة المتزامنة، بمساعدة مراجع لغوية منتقاة بعناية، تقليص خطر تحيز التركيب بشكل كبير في سياقات التقييمات المتقاطعة ثقافياً (انظر إلى المثال 13)، وتسمح بعدم مركزية لغوية وثقافية بدرجة أكبر بكثير من تكييفات الاختبارات المتعاقبة.

التطوير المتزامن لاختبار تركيزي Concentration test (المثال14):

على نقيض الـ d2، تم تصميم سلسلة من الاختبارات المتعاقبة PTS، غيتلر وتانزر، 1995/1998 بشكل مسبق للاستخدام في الإعدادات المتعددة الثقافات واللغات. وهو يتألف من 9 اختبارات فرعية مع تعقيد معرفي متزايد، يتراوح بين الاختبارات الثانوية التي تتطلب فقط سرعة حسية وحركية، إلى الاختبارات الثانوية التي تستلزم عناصر من الذكاء السلس. مثل "الدمية الروسية". تتطلب المهمة المعرفية لكل من صيغ الاختبار الثانوية، جزءاً من الاختبار الثانوي التالي. وهكذا، تنتج بنية شبه مبسطة. يقوم هذا التصميم أيضاً بتسهيل تعليمات الاختبار، لأن الفرق بين اختبارين ثانويين، الذي يحتاج تفسيراً هو فقط.

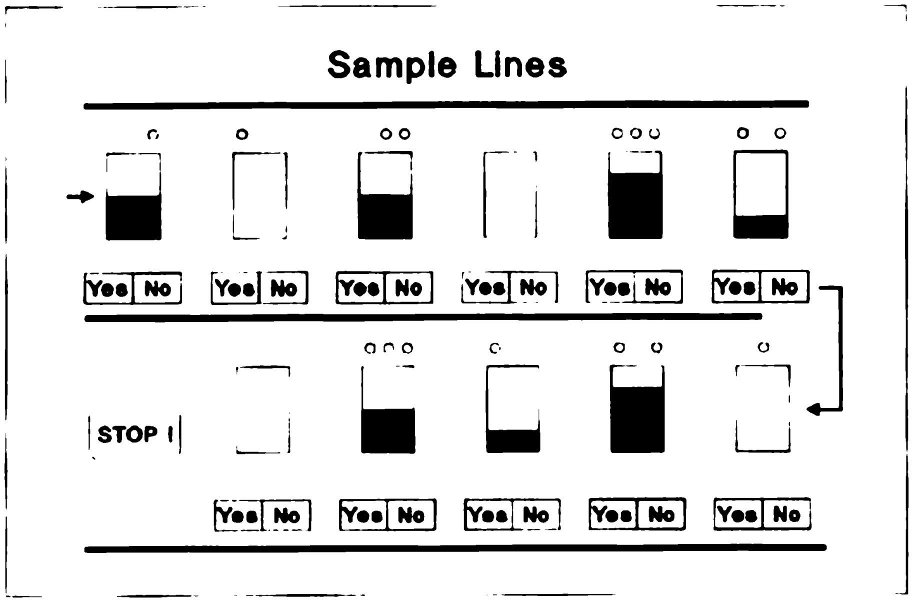
من أجل تجنب تحيز عرقي في مادة الاختبار الأصلية، تتألف أسئلة الاختبارات الثانوية التسعة من عناصر هندسية بسيطة (نقاط، دوائر، إطارات، شرائط، مناطق مظللة بالرمادي) لا بد أنها مألوفة للأشخاص المتعلمين في كل الثقافات (توجيه 6)، بطريقة مماثلة، إن خيارات الإجابات المعروضة ("0، 1، 2" "1، 2، 3"، أو " نعم/ لا") تحت كل سؤال، يمكن ترجمتها بسهولة إلى كل لغة (توجيه 5). بالإضافة إلى ذلك، فإن "معرفة الكيفية" المحصلة في تكييف d2 المتعدد الثقافة واللغة (انظر إلى الأمثلة 10-12) تم استخدامها لإعادة تصميم مواد الـ PTS (التوجيهات 13-14) كما هو موضح في الشكل رقم 10 .6، تم استخدام المساعدات البصرية المتعلقة بالمهمة وتعليم تصحيح الإجابات. للموازنة بين استعمال يد معينة (أثناء الحل) وتوجيهات الكتابة في لغة المرء الأصلية (توجيه 1)، على المختبرين أن يعملوا خلال الاختبار بشكل "متعرج " كما هو موضح في الشكل 10 .7. أخيراً، لضمان أنه حتى المختبرين ذوي التعليم الضئيل يمكن أن يفهموا المهمة المطلوبة (توجيهات 10.4)، يمكن أن تعطى التعليمات بذلك عبر "قصة غلاف واقعية". إن قصة الغلاف المعاد صياغته للاختبار الثانوي (انظر أيضاً إلى الشكل 10 .6) هو كالتالي:

انظر إلى الإطارات [أو"النوافذ"] يمكن أن تكون فارغة أو معبأة من واحد إلى ثلاثة شرائط [أو"كتل مستطيلة"] رمادية. في الإطار الأول، يمكنك أن تحصي عدد الشرائط الرمادية. في الإطارات الأخرى، يجب أن نقدرهم من أعلى المنطقة المظللة بالرمادي. تشير الدوائر فوق كل إطار إلى عدد الشرائط التي يجب أن تكون في الإطار. إن مهمتك هي أن تقرر [نعم/ لا] لكل إطار إذا كانت الحالة كما هي الآن..



الشكل رقم 6.10

مثال لبند أُستخدم في اختبار C الثانوي السلاسل اختبار تدريجي (PTC)



الشكل رقم 7.10

مثال لبنود استخدمت في اختبار C الثانوي لسلاسل اختبار تدريجي (PTS)

خاتمة

إن تطوير الاختبارات للاستخدام في الثقافات واللغات المتعددة يستلزم الدمج الناجح لعدد من الاعتبارات النفسية، التقاطعات الثقافية واللغوية، من أجل ضمان صدق التقاطع الثقافي واللغوي للتركيب الضمني، وتصميم الاختبار لقياسه، وإدارة الاختبار والنتائج التي تم تحصيلها من درجات الاختبار. بافتراض القصور لإجراءات تكييف الاختبار المتعاقبة حتى في حالة الاختبارات السهلة نسبياً مثل d2، نوصي بمعالجة متزامنة لتطوير الاختبارات المتقاطعة ثقافياً ولغوياً؛ لأنها تسمح بالحد الأقصى من عدم المركزية اللغوية/ الثقافية. ولكن هذه الطريقة يمكن أن تضمن أدوات متقاطعة ثقافياً ولغوياً فقط إذا تحققت اللوازم التالية:

1- برهان لتكافؤ التركيب. من المتطلبات d2 الأساسية لتطوير أداة صادقة للاستعمال في اللغات والثقافات المتعددة هي قابلة التعميم للتراكيب الضمنية عبر كل مراجع الثقافات واللغات (توجيه2). بالإضافة إلى البينة التي تم جمعها من أدبيات الثقافة (الأصلية)، وعلم النفس عبر الثقافات (المقارن) وعلم الأعراق البشرية وعلم اللغويات، فإن كل من البيانات التجريبية والحكمية يجب أيضاً أن تجمع خلال عملية تطوير/ تكييف الاختبار (توجيهات 9.7).

2- تصنيف تكافؤ التقاطع الثقافي واللغوي. مطلب أساسي آخر هو تصنيف تكافؤ التقاطع الثقافي واللغوي. يجب أن يحدد هذا التصنيف تحت أية شروط يمكن لأنواع معينة من النتائج (مثال: تحميل العامل، علامات الاختبار) التي تم تحصيلها من نسخ لغوية مختلفة أو من مجموعة ثقافية (توجيهات 12-20-22). تم تقديم اقتراح شامل نسبياً لمثل هذا التصنيف من قبل فان دي فيفر ولونغ (1997a ،1997 b).

3- تصنيف التحيز المتعدد الثقافة/ المتعدد اللغة. مطلب أساسي إضافي هو تصنيف شامل لمصادر مختلفة من تحيز التركيب والقياس (يتضمن الطريقة،

الإدارة وتحيز السؤال). في تطبيقات الاختبارات المتقاطعة ثقافياً ولغوياً. يجب أن يحدد هذا التصنيف أي من أنواع التكافؤ من المرجع أن تتأثر، وبأية أشكال من التحيز. ويجب أن تؤكد بمجموعة كبيرة من التوضيحات التجريبية (التوجيهات 13-14). مع التصنيفات المقترحة من فان دي فيفر ولونغ (b 1997، 1997a) أو فان دي فيفر وتانزر (1997)، يعد توجيهات تكييف الاختبار لـ ITC مع توضيحاتهم المنطقية والتجريبية خطوة تمهيدية نحو هذا التصنيف الشامل.

4- تصنيفات الحلول. إن تركيب "كتاب طهي" يحدد وصفات لكل المصادر التي ينشأ منها التحيز في التقييمات المتعددة الثقافة واللغة هو، على الأغلب، مشروع لا أمل منه. رغم ذلك، يحتاج التقدم المستقبلي في الاختبارات المتعددة الثقافات واللغات، تصنيفاً للحلول (توجيهات 1، 13) يستند على تصنيف متقاطع لأنواع مختلفة من التكافؤ ومصادر مختلفة من التحيز. إن تصنيف الحلول الذي اقترح من قبل فان دي فيفر وتانزر (1997) يعد خطوة أولى في هذا الاتجاه. ولكن من أجل تقديم رؤية أكثر عالمية، يجب أيضاً أن يتم دمج البيانات من نظرية القياس وعلم النفس عبر الثقافات (غرينفيلد).

5- البنية المنهجية لتقصي وتقييم ومعالجة التحيز. بدون منهجية ثابتة لتقصي مصادر التكافؤ، فإن التصنيفات المذكورة سابقاً للتحيز والحلول، لا يمكن أن تحول إلى تطبيقات عملية (التوجيهات 7-9، 11، 13). يجب أن تحدد البنية المنهجية المطلوبة أي الإستراتيجيات المؤثرة من أجل تقصي مصادر التحيز، وأيضاً لتقييم تأثير المقاييس المقابلة المأخوذة. "استراتيجيات صندوق الأدوات" هذه، يجب أن تتضمن إجراءات حكمية. وتقنيات قياسية (مطبقة في برامج كمبيوتر تتوفر بسهولة ومرنة مع المستخدم)، بالإضافة إلى تصاميم تجريبية (مثال، غيتلر وتانزر، 1998) يجب أن تشمل أيضاً توجيهات على مجموعة المعلومات الإضافية (مثال، تسجيل وقت العمل في سرعة الاختبار الذاتية أو

استخدام تقنيات" التفكير بصوت عالي") لأنها تعطي رؤية أعمق إلى مميزات ثقافية محتملة لعمليات تحفيزية ومعرفية (التوجيه 22).

6- الأرشيف الموثق. إن إمكانية الدخول إلى أرشيف كبير ومحفوظ بشكل جيد يقدم مستندات نصية مفصلة على تكييف وتطوير الاختبارات المتعددة الثقافة واللغة (التوجيهات 4-6، 11-10، 19-17،22) سوف يساعد على التحديث المستمر للتصنيفات المذكورة سابقاً للتحيز وتصنيف الحلول، وعلى تقييم جدواها على أساس بينات جديدة. في حال كان هذا الأرشيف محفوظاً بشكل جيد، فسيكون بمقدور مطوري الاختبارات الاسترشاد بتصنيف الحلول لتحديث المقاييس المقابلة المؤثرة (التوجيه 13). إن الدراسة الحديثة لهامبلتون، ولي وسيرسي (2002) التي بين فيها الباحثون العديد من الأخطاء الشائعة التي وجدت في مطبوعات تكييف الاختبارات، هي خطوة أولية نحو الأرشيف المقترح.

7- المشروع التكميلي لمطوري الاختبارات. إن مطوري الاختبارات المتعددة الثقافة واللغة ليسوا فقط بحاجة لأن تكون لديهم الكفاءة في اللغات المتعددة وخبرة في علم النفس السائد (بما فيها المعرفة بالتركيب ومقياسه)، وعلم القياس وتقنيات تركيب الاختبار، ولكنهم بحاجة أيضاً إلى الكفاءة في علم اللغات والثقافة المتقاطعة والثقافة (أي الأصلية) النفسية، بالإضافة إلى العلم بالميزات الثقافية (مثال، الموضوعات المحظورة، غودوين ولي، 1994). في كل المراجع الثقافية المعنية. من الواضح أنه لا يمكن وجود شخص يستطيع جمع كل هذه المجالات المختلفة من الخبرات: لذا، يجب أن يتم التركيز على كفاءة إجرائية تتعلق بإعداد وإدارة ناجحة للجنة عمل متعددة الثقافات واللغات ذات خبرة من الدراسات المكملة.

8- مؤهلات مستخدمي الاختبارات. بالإضافة إلى الخبرة في الإدارة، والدرجات، وترجمة اختبار معين، على مستخدمي الاختبارات أن يملكوا كفاءة عامة في إجراء التقييمات في السياق المتعدد الثقافة واللغة (التوجيهات 14، 15، 18).

وهذا يشمل حساً ثقافياً و كفاءة في التواصل عبر الثقافات (اسانت وغوديكونست، 1989، كوهين، 1987، لانديس و بهجت، 1996، سشنيلر، 1995، ويرلي، 1995).

واضح أن كل هذه المتطلبات تتجاوز المقدرات الفردية لمطوري ومستخدمي الاختبارات؛ لذلك، فإن تطوير وتأسيس البرامج المؤهلة لمطوري ومستخدمي الاختبارات، التي تغطي هذه المتطلبات، يجب أن يكون موضوعاً رئيساً للمنظمات العلمية والحرفية المهتمة بالاختبار النفسي والتربوي.

*****

المراجع

Asante, M. K., & Gudykunst, W. B. (Eds.). (1989). *Handbook of international and intercultural communication*. London: Sage.

Bartram, D., & Coyne, I. (1998). *The ITC/EFPA survey on testing and test use in countries world-wide. Narrative report* (Research Report). Hull, England: University of Hull, Psychology Department.

Bond, M. H. (1990). *Beyond the Chinese face. Insights from psychology*. Hong Kong: Oxford University Press.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brickenkamp, R. (1962). *Aufmerksamkeits-Belastungstest (Test d2)* [The d2 test of attention] (1st ed.). Goettingen, Germany: Hogrefe.

Brickenkamp, R., & Zillmer, E. (1998). *d2. Test of attention*. Seattle: Hogrefe & Huber.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 389–444). Boston: Allyn & Bacon.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Newbury Park, CA: Sage.

Broer, T. (1996). *Rasch-homogene Leistungstests (3DW, WMT) im Kulturvergleich Chile-Österreich. Erstellung einer spanischen Version einer Testbatterie und deren interkulturelle Validierung in Chile* [Cross-cultural comparison of the Rasch-calibrated tests 3DC and VMT between Chile-Austria and the development of a Spanish version of the test battery]. Unpublished master's thesis, University of Vienna, Austria.

Cheung, F. M. (2004). Use of western and indigenously-developed personality tests in Asia. *Applied Psychology: An International Review, 53*(2), 173–191.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Chang, J. P. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology, 27,* 181–199.

Cohen, R. (1987). International communication: An intercultural approach. *Cooperation and Conflict, 22,* 63–80.

Conklin, J. (1987). Hypertext: An introduction and survey. *IEEE Computer, 20,* 17–41.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6,* 397–416.

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest* [The Viennese Matrices Test. A Rasch-calibrated nonverbal intelligence test]. Weinheim, Germany: Beltz Test.

Gao, G. (1998). "Don't take my word for it."—Understanding Chinese speaking practices. *International Journal of Intercultural Relations, 22,* 163–186.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Gittler, G. (1990). *3DW. Dreidimensionaler Würfeltest. Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Theoretische Grundlagen und Manual* [The Three-Dimensional Cube Test, 3DC. A Rasch-calibrated spatial ability test. Theoretical background and test manual]. Weinheim, Germany: Beltz Test.

Gittler, G., & Tanzer, N. K. (1990/1998). *The Progressive Test Series (PTS)*. Unpublished test, University of Vienna and University of Graz, Austria.

Gittler, G., & Tanzer, N. K. (1998, August). *Establishing cross-cultural equivalence of item complexity using the linear logistic test model (LLTM)*. Paper presented at the 24th International Congress of Applied Psychology, San Francisco.

Goodwin, R., & Lee, I. (1994). Taboo topics among Chinese and English friends. A cross-cultural comparison. *Journal of Cross-Cultural Psychology, 25,* 325–338.

Greenfield, P. M. (1997). You can't take it with you. Why ability tests don't cross cultures. *American Psychologist, 52,* 1115–1124.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment (Bulletin of the International Test Commission), 10,* 229–244.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., Li, S., & Sireci, S. (2003). *Pitfalls and obstacles in the test adaptation process: A meta-analysis* (Center for Educational Assessment Research Report No. 489). Amherst, MA: University of Massachusetts, School of Education.

Håseth, K. J. (1996). The Norwegian adaptation of the State–Trait Anger Expression Inventory. In C. D. Spielberger & I. Sarason (Eds.), *Stress and emotion* (Vol. 16, pp. 83–106). Washington: Taylor & Francis.

Hodapp, V., Tanzer, N. K., Maier, E. R., & Pestemer, I. A., & Korunka, C. (in press). The German adaptation of the Job Stress Survey: A multi-study validation in different occupational settings. In C. D. Spielberger & I. G. Sarason (Eds.), *Stress and emotion* (Vol. 17). Washington, DC: Taylor & Francis.

Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children and youth: An empirical study. *International Journal of Psychology, 26,* 329–244.

Landis, D., & Bhagat, R. S. (Eds.). (1996). *Handbook of intercultural training* (2nd ed.). London: Sage.

Marsh, H. W. (1988). *Self-Description-Questionnaire I. SDQ–I manual and research monograph.* San Antonio, TX: Psychological Corporation.

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 131–198). Orlando, FL: Academic Press.

Marsh, H. W., Craven, R., & Debus, R. (1999). Separation of competency and affect components of multiple dimensions of academic self-concept: A developmental perspective. *Merril–Palmer Quarterly, 45,* 567–601.

Martini, D. R., Strayhorn, J. M., & Puig-Antich, J. (1990). A symptom self-report measure for preschool children. *Journal of the American Academy of Child and Adolescent Psychiatry, 29,* 594–600.

McFadden, J. (Eds.). (1993). *Transcultural counseling: Bilateral and international perspectives.* Alexandria, VA: American Counseling Association.

Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin, 112,* 179–204.

Moosbrugger, H., & Oehlschlägel, J. (1996). *FAIR. Frankfurter Aufmerksamkeitstest* [Frankfurt attention test]. Bern, Switzerland: Huber.

Oakland, T. (1997). Test use among school psychologists: Past, current, and emerging practices. *European Journal of Psychological Assessment, 13,* 2–9.

Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied Psychology: An International Review, 53*(2), 157–172.

Oakland, T., & Hu, S. (1992). The top 10 tests used with children and youth worldwide. *Bulletin of the International Test Commission, 19,* 99–120.

Paniagua, F. A. (1994). *Assessing and treating culturally diverse clients: A practical guide.* Thousand Oaks, CA: Sage.

Piswanger, K. (1975). *Interkulturelle Vergleiche mit dem Matrizentest von Formann* [Cross-cultural comparisons with Formann's Matrices Test]. Unpublished doctoral dissertation, University of Vienna, Vienna, Austria.

Ponterotto, J. G., Casas, J. M., Suzuki, L. A., & Alexander, C. M. (Eds.). (1995). *Handbook of multicultural counseling.* Thousand Oaks, CA: Sage.

Rosenzweig, S. (1977). *Manual for the children's form of the Rosenzweig Picture-Frustration (P-F) Study.* St. Louis, MO: Rana House.

Schneller, R. (1989). Intercultural and intrapersonal processes and factors of misunderstanding: Implications for multicultural training. *International Journal of Intercultural Relations, 13,* 465–484.

Schwenkmezger, P., Hodapp, V., & Spielberger, C. D. (1992). *Das State-Trait-Ärgerausdrucks-Inventar STAXI* [The German adaptation of the State-Trait Anger Expressions Inventory]. Bern, Switzerland: Huber.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–166.

Spielberger, C. D. (1988). *State–Trait Anger Expression Inventory research edition. Professional manual.* Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D., & Comunian, A. L. (1992). *STAXI. State–Trait Anger Expression Inventory. Versione e adattamento italiano a curi di Anna Laura Comunian. Manuale* [Test manual of the Italian version of the State–Trait Anger Expression Inventory]: Firenze, Italy: Organizzazioni Speziali.

Spielberger, C. D., & Reheiser, E. C. (1994). Job stress in university, corporate, and military personnel. *International Journal of Stress Management, 1,* 19–31.

Tanzer, N. K. (1995). Cross-cultural validity of Likert-type scales: Perfect matching factor structures and still biased? *European Journal of Psychological Assessment, 11,* 194–201.

Tanzer, N. K. (1998). *Assessment of domain specificity in school-related Likert-type inventories: Conceptual issues, psychometric approaches, and cross-cultural evidence.* Unpublished habilitation monograph. Graz, Austria: University of Graz.

Tanzer, N. K., Ellis, B. B., Gittler, G., & Broer, T. (1996, August). *The use of collateral information in establishing the cross-cultural validity of tests.* Paper presented at the 26th International Congress of Psychology, Montreal, Canada.

Tanzer, N. K., Ellis, B. B., Zhang, H.-C., Sim, C. Q. E., Broer, T., & Gittler, G. (1997, July). *Cross-cultural decentering of test instructions in a letter-cancellation test: A field test of the ITC Guidelines for Test Adapations.* Paper presented at the 5th European Congress of Psychology, Dublin, Ireland.

Tanzer, N. K., Gittler, G. & Ellis, B. B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment, 11,* 170–183.

Tanzer, N. K., Gittler, G., & Sim, C. Q. E. (1994). A cross-cultural comparison of a Rasch calibrated spatial ability test between Austrian and Singaporean adolescents. In A. Bouvy, F. J. R. van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 96–110). Lisse, Netherlands: Swets.

Tanzer, N. K., & Sim, C. Q. E. (1991). *Self-concept and achievement attributions. A study of Singaporean primary school students* (Research Rep. No. 1991/5). Graz, Austria: University of Graz, Department of Psychology.

Tanzer, N. K., Sim, C. Q. E., & Marsh, H. W. (1992). Using personality and attitude inventories over cultures: Theoretical considerations and empirical findings. *Bulletin of the International Test Commission, 19,* 151–171.

Tanzer, N. K., Sim, C. Q. E.,& Marsh, H. W. (1997). *Where cross-cultural and differential psychology meet: Competence/task-easiness and interest/eagerness as subcomponents of academic self-concept* (Research Rep. No.1997/1). Graz, Austria: University of Graz, Department of Psychology.

Tanzer, N. K., Sim, C. Q. E., & Spielberger, C. D. (1996). Experience and expression of anger in a Chinese society. The case of Singapore. In C. D. Spielberger & I. Sarason (Eds.), *Stress and emotion* (Vol. 16, pp. 51–65). Washington, DC: Taylor & Francis.

Toubiana, Y. (1994). *Pictorial evaluation of test reactions (PETER).* Petach-Tikva, Israel: PETER.

Unterholzner, B. (1997). *Validierung einer italienischen Form des Prüfungsstressinventars (PSI) von Tanzer* [Validation of an Italian adaptation of Tanzer's Examination Stress Inventory]. Unpublished master's thesis, University of Graz, Austria.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 1, pp. 257–300). Chicago: Allyn & Bacon.

van de Vijver, F. J. R., & Leung, K. (1997b). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 4,* 263–279.

Van Haaften, E. H., & van de Vijver, F. J. R. (1996). Psychological consequences of environmental degradation. *Journal of Health Psychology, 1,* 411–429.

Wehrly, B. (1995). *Pathways to multicultural counseling competence: A developmental journey.* Pacific Grove, CA: Brooks/Cole.

Yang, K.-S., & Bond, M. H. (1990). Exploring implicit personality theories with indigenous or imported constructs: The Chinese case. *Journal of Personality and Social Psychology, 58,* 1087–1095.

Zhang, J. X., & Bond, M. H. (1998). Personality and filial piety among college students in two Chinese societies. The added value of indigenous constructs. *Journal of Cross-Cultural Psychology, 29,* 402–417.

11

القياس النفسي للتكييف: تقييم تعادل القياس عبر اللغات والثقافات

فيتز دراسكو وتاهيرا برويسن
جامعة الينوي

إننا نجد أن معاهدة التجارة الحرة في أمريكا الشمالية (NAFTA) وتشابك الوحدة الأوربية هما مجرد مؤشرين على الاتجاه المتزايد نحو العولمة في الأسواق الدولية. وكمؤشر على وتيرة التغير نجد أن الصادرات التجارية الدولية ارتفعت من (7.382) بليون دولار في سنة 1985 إلى /1.1/ تريليون دولار عام 1996 (باتش 1998). وبالإضافة إلى تشكيل دول لاتحادات اقتصادية كبرى، فقد أصبح شيئاً اعتيادياً لشركات كبيرة في دول مختلفة مثل شركة "كرايسلر" و"ديملربنز" أن تندمج لتجعل نشاطاتها التجارية متكاملة. وكنتيجة لعولمة أسواق العالم تصبح الحاجة لفهم الفروق الحضارية بين الناس المتباينين في المشارب والأصول الحضارية أكثر وضوحاً.

وقد كان من المعالم المهمة في التمازج بين الثقافات نشر كتاب "عواقب الثقافة" في عام 1980 لمؤلفه "جرته هوفستيد" الذي دوّن أبعاداً متعددة ومهمة تختلف بموجبها الثقافات. فبالتحليل العاملي لما يزيد عن 116 ألف استبانه أخذت من موظفي شركة "IBM" في "40" بلداً مختلفاً، وجد هوفستيد أن الثقافات تختلف في كلٍّ من مسافات النفوذ Power distance، وتجنب اللامحدودية Uncertainty، والرجولة Masculinity، والفردية Individualism. فمسافات النفوذ تشير إلى

القابلية التي ترى بها المسافات الواسعة بين المستويات الأعلى والأدنى في الهرم الاجتماعي. أما تجنب عدم الوضوح في المواقف فيعكس تجنب المواقف التي تكون النتيجة فيها غير واضحة. والرجولة تعني نظرة أعضاء المجتمع الحضاري إلى أنشطة قيم تعتبر عرفياً أكثر رجولة في ثقافتهم، ولكن البعد المدروس بتمحيص أكثر من غيره بكثير فهو البعد الذي يتناول الفردية والجمعية (الجماعية).

وقد سجل ترياندس (1990-1994-1995) بعض الاختلافات القوية، الواضحة التي ينظر فيها الناس، في بعد الثقافات الفردية والجماعية في العالم؛ في بعض الثقافات الفردية مثل الولايات المتحدة الأمريكية وكندا وكثير من دول غرب أوربا، يركز الأفراد على الذات. فالأفراد يؤكدون على الحاجات والرغبات لدى الفرد قبل حاجات الجماعة ورغباتها. ويفترض بالأفراد أن يكونوا مستقلين بأنفسهم، ويتقرر السلوك على الأغلب بالأهداف الفردية. وكنتيجة نجد أن الشعور الجماعي ليس له تأثير كبير على الأفراد. ومن الناحية الأخرى نجد أمريكا الجنوبية وأمريكا الوسطى وآسيا وأجزاء عديدة في العالم المتطور جماعية النزعة. ومن وجهة نظر ترياندس (1990) يؤكد أولو النزعة الجماعية على وجهات النظر والحاجات والأهداف المشتركة أكثر من حاجات الذات وهنا تنحو رغبات الفرد لأن تكون تابعة للحاجات المشتركة. وتقرر نمط السلوك بشكل كبير بالعرف الاجتماعي والواجبات أكثر من الرغبات الفردية، كما أن التناسق الاجتماعي المشترك يعد ذا قيمة عالية.

وعلى الرغم من أن الباحثين في علم النفس التنظيمي في الصناعة، وباحثي إدارة الموارد البشرية يدركون أن الفروقات الثقافية لا يمكن تجاهلها في المنظمات متعددة الجنسيات، فإن البحث الإجرائي في هذه الموضوعات قد تقدم بشكل بطيء بسبب المشكلات النظرية والعملية العديدة. فالحصول على فرصة لإجراء البحث في المنظمات متعددة الجنسيات باستخدام التسهيلات الدولية أكثر صعوبة من الحصول عليها في مجالات الأعمال الأمريكية. وبعد الحصول على الإذن بإجراء

البحث يتبين أن التكلفة تكون عالية بشكل لا يشجع على إجراء البحث. إلا أنه قد تكون أشق عقبة يتوجب تخطيها في محاولة إجراء البحث الذي يتناول موضوع دراسة الثقافات سوية بشكل شامل باستخدام الأساليب السليمة، تكمن عند مرحلة تكييف الأدوات المستخدمة، فقد قيل عن التكييف إنه: "قد يكون أكثر أنواع الأحداث تعقيداً بين الأحداث التي ظهرت حتى الآن منذ تطور الخليقة". (ريتشاردز 1953 ص250)؛ وعلى الرغم من أن عبارة ريتشاردز لا شك قد تكون مبالغاً فيها بالطبع، إلا أن هذا الفصل، بل هذا الكتاب بأكمله هو عبارة عن تقرير حول التحديات الكامنة في إجراءات التكيف والترجمة.

هناك العديد من الخطوات التي يحتاج الباحثون لاتخاذها لضمان التعادل في النوعية والقياس بين الأدوات التي يتبنونها لإجراء بحثهم، وهذه الخطوات شديدة الأهمية، إذ بدونها قد تهدر كميات هائلة من وقت البحث ومن الجهد المبذول في سبيله ومن الأموال دون طائل. يبدأ مسار تكييف الأداة المستخدمة بإجراء استبيان شامل يوضع أساساً في العادة بلغة معينة من اللغات المصدر (اللغة المترجم منها)، وعند إجراء البحث في المقارنة بين الثقافات كثيراً ما يكون ضرورياً أن تجرى تكييفات للأداة في لغة أو أكثر من اللغات المتلقية للترجمة، أو اللغات التي يجري التركيز عليها. والشكل الأفضل يكون باستخدام مجموعتين من المترجمين لهذا الغرض، يقوم أفراد المجموعة الأولى من المترجمين بشكل إفرادي بترجمة الاستبيان من لغة المصدر إلى اللغة المتلقية وبعد ذلك يقوم أفراد هذه المجموعة بالاجتماع وتسوية الاختلافات في ترجماتهم.

وبعد أن يتم الاتفاق على الشكل النهائي يعطى الإستبيان لكل مترجم من الفئة الثانية فيقوم هؤلاء المترجمون بدورهم، كل بمفرده، بإعادة ترجمته من اللغة المتلقية إلى اللغة الأصل، وبعد أن يسوي هؤلاء المترجمون بدورهم اختلافاتهم يقارن

الاستبيان الأصلي مع الورقة المعاد ترجمتها وفي نهاية هذه الإجراءات الطويلة المعقدة يكون من دواعي الارتياح والطمأنينة أن تظهر النسخة الأخيرة للترجمة متشابهة بشكل عالٍ مع النسخة الأصلية من الاستبيان.

إلا أنه كثيراً ما يكون هناك تفاوت في أحوال عديدة بين النسخة الأصلية والنسخة المعاد ترجمتها؛ فمثلاً من خلال العديد من المحاولات التي تبذل للتكييف واجهتنا مراراً صعوبات في ترجمة التعابير الأمريكية المصطلح عليها اجتماعياً إذ عندما نصف عملا على أنه «مريح» يفهم الأمريكيون ذلك إلا أنه في لغة "المراثي Marathi" ترجمت هذه العبارة بكلمة "يؤدي بشكل سهل" وكذلك يمكن لنا أن نرى أن وصف عبارة (waste of time) أي «هدر الوقت» في اللغة الإنكليزية قد يترجم على أنه (وقت زملاء العمل المهدور).

وبالإضافة إلى هذه التعابير المتعارف عليها يمكن أن تقابل الصعوبات أيضاً عند ترجمة قطعة بسيطة نسبياً من النثر، واللغة الإنكليزية لغة فيها الكثير من المترادفات. وأحياناً تصاغ المقاييس بوضع السؤال ذاته باستخدام أشكال متعددة من الكلمات أو العبارات المتشابهة، وحيث إن المترادفات كثيراً ما تنشأ لغوياً "عندما يكون المفهوم ذا أهمية في ثقافة معينة (فمثلاً لدى الإسكيمو أكثر من مئة كلمة لوصف الثلج، فإن الثقافات الأخرى قد لا يكون لديها إلا وصف واحد للكلمة). وكمثال في بحثنا الذي يتناول التوتر بشكل عام، يسأل مقياس (SIG) الذي وضعه سميث وساديمان ومكراري في عام (1962) المستجيبين أن يذكروا إذا كان عملهم "مزحوماً" أو "داعياً للإزعاج"، إلا أنه في لغة "المراثي" لا تجد سوى كلمة "مزعج" كأقرب وصف يمكن ترجمة المفهوم إليه. وعلى ذلك، عند استخدامنا لمقياس (SIG) بشكله الكامل كان لا بد لنا أن نكتفي باستخدام كلمتي "مزعج" و"مزعج جداً" عند ترجمة هاتين الكلمتين. وعند حدوث تباينات وتغيرات في المدلول في الترجمة بين اللغة مصدر النص واللغة المتلقية عندئذٍ يجب علينا في هذه الحالة أن نكرر إجراء

الترجمة إلى أن يتم الحصول على الامتزاج السليم. وطبعاً "لا تزال هناك معضلة أخرى في سر التكييف وهي تقرير ما يكون عليه الامتزاج السليم (الوافي)". وعلى الرغم من وجود درجة عالية من التشابه بين النص الأصلي والترجمات المعادة ترجمتها من الاستبيان إلا أن ذلك لا يكفل التعادل بين النص في اللغة الأصلية والنص في اللغة المترجم إليها. فمثلاً يمكن للترجمة المشوشة أن تعاد إلى ما يشبه النص الأصلي إذا تمكن الذين يعيدون الترجمة من التكهن بما يعنيه المترجمون الأولون. وهناك مشكلة أقوى سببها أن الأفراد الذين يتقنون لغتين لا يشبهون في أي من اللغتين أولئك الذين يتقون لغة واحدة فقط في أي من الجانبين (لاندر وايرفنغ وهورومبز 1960). وأهمية هذه النقطة أكثر من مجرد أكاديمية، فعندما درسنا نوعية ترجمة لمقياس يقيم الرضى بالعمل باستخدام ذوي اللغتين (هولين ودراتكو وكريوكار 1982) لم نجد الكثير من دلائل عدم التكافؤ بين اللغتين الإنكليزية والإسبانية (مجرد 3 بنود من أصل 72 بند كانت التي ظهر فيها الاختلاف في القياس بين اللغات) ولكن فيما بعد عندما قارنا الشكلين ذاتهما في مقياس "الرضى بالعمل" بتطبيقه على متكلمي اللغة الإنكليزية وحدها ومتكلمي اللغة الإسبانية وحدها (دراسكو وهولين 1988) وجدنا كثيراً من الاختلافات عبر اللغتين (حوالي ثلث بند المقياس). وبالرغم من متابعة مسار التكيف المذكور أعلاه بدقة هناك على الأقل ثلاثة أسباب لكون تحليل خواص القياسات لبند المقياس قد تكشف عن اختلافات عبر اللغات. فالسبب الأول والأكثر وضوحاً هو أن البنود لم تكن مترجمة بشكل سليم. فمثلاً الفرق بين كلمة حقل وكلمة مزرعة تربية أبقار أمريكية بيّن، واختيار كلمة في اللغة التي تجري الترجمة إليها تقارب كلمة مزرعة الأبقار تماماً شيء صعب جداً. ثانياً قد تكون بعض المفاهيم المألوفة في إحدى الثقافتين صعبة أو مستحيلة الفهم على أعضاء ثقافة أخرى. فمثلاً عبارة "ما حك جلدك مثل ظفرك" (Do your own thing) وعبارة: "لا يعيش الإنسان إلا مرة واحدة" (You only live once) تضعان التركيز على الذات التي تفهم، بل وينادى بها أيضاً، في البلدان ذات

الثقافات فردية الاتجاه، إلا أن مثل هذا التجاهل للأصدقاء والعائلة يمكن أن لا يفهم بسهولة من قبل أفراد ثقافات تركز على المجموعة أكثر من تركيزها على الفرد. وأخيراً يعتقد أحياناً أن الأفراد القادمين من ثقافات مختلفة يميلون إلى استخدام مقاييس الإجابة بشكل مختلف. وبشكل محدد ينحو بعض الأفراد إلى تجنب استخدام الشكل الأقصى لمقياس الإجابة (مثلاً الاختيار بين /1/ و/7/ من مقياس مثل مقياس الإجابة "ليكرت" المكون من 7 بنود)، بينما يميل أفراد ثقافات أخرى إلى إعطاء تقييمات مبالغ فيها (هوي وتزياندليز 1989، ترياندرلز 1972) فمثلاً وجد ماران وكامب وماران (1992) أن متكلمي اللغات الإسبانية أكثر ميلاً لاستخدام إجابات متطرفة وأن يتفقوا مع التقارير أكثر من الشعوب غير الإسبانية فهم يعكسون جزئياً فروقاً في الأعراف الاجتماعية المتعلقة بالموافقة.

تجعل هذه المشكلة المرء يرى بشكل جلي أن مقارنة إجابات الأشخاص الذين يتكلمون لغات مختلفة وينشؤون في ثقافات مختلفة أمر عسير جداً - إذ كيف يمكن للباحثين أن يقروا إجرائياً ما إذا كانت البنود والمقاييس المعطاة لمثل هذه المجموعات المتباينة من الناس تقيس بشكل متعادل فعلياً، هدف هذه الفصل هو توضيح منطلق واحد لمثل هذه التحليلات فنحن نستخدم نظرية الإجابة وفق البنود (IRT) لدراسة ما إذا كانت العلاقة بين احتمال اعتماد بند والخاصية الخفية الكامنة التي يقيسها المقياس هي متماثلة لدى المجموعات كافة.

نظرية إجابة البنود Item Response Theory

نقدم في هذا الفصل وصفاً موجزاً لهذه النظرية (IRT) يقدم هاملتون وسواميناثان (1985) وبيكر (1992) وهولين ودرازكو (1983) تفصيلات أكثر توسعاً حولها.

لنفرض أن U1 وU2 وUn... تصف عدداً (ن) (n) من البنود على مقياس ما. لتبسيط الأمور لا نعتبر إلا نماذج الإجابات ذات الدرجات المأخوذة ثنائياً، هنا (Ui=1) من أجل إجابة إيجابية لبند موضوع بشكل إيجابي (مثلاً إجابة بنعم لبند

يسأل: "هل يبعث عملك على الرضى في نفسك؟") أو جواب بالنفي لبند مصوغ بشكل سلبي مثلاً: جواب بالنفي على سؤال يقول: "هل عملك جميل؟" المتغير الثنائي Ui=0 لجواب بالنفي مصاغ إيجابياً (مثلاً جواب كلا لسؤال يسأل: (هل عملك رائع؟") أو جواب إيجابي لسؤال سلبي الصيغة: ("هل عملك يقودك للإحباط؟") وقد أصبحت النماذج المتعددة الأبعاد التي تسمح بتصنيف الإجابات في زمر منظمة (مثلاً نموذج سمجيما 1969) متدرج الإجابات، أو الزمر الاسمية (مثلاً النموذج الاسمي الذي وضعه "بوك" (1972) شائعة ومرغوبة على نطاق واسع (عدد آذار 1995 من مجلة القياسات النفسية التطبيقية الذي كان مكرساً للنماذج متعددة الأبعاد) ولكن ذلك خارج نطاق بحثنا الآن.

وكذلك لا ننظر إلا إلى نماذج الإجابة للبند التي تكون أحادية الأبعاد (ولكن يمكن ملاحظة أن عدد كانون الأول 1996 من مجلة القياس النفسي التطبيقي كان مكرساً لنماذج متعددة الأبعاد من الإجابة للبنود). بالنسبة للنماذج أحادية الأبعاد تستخدم قيمة - (ثيتّا) المقياسية في كثير من الأحيان لتعني الخاصية الخفية الكامنة المقيّمة موجب (n)"ن" على المقياس، ووظيفة إجابة البند /IRF/ التي تسمى أحياناً المنحنى المميز للبند هي أساسية في هذه النظرية، فهي تعطي احتمال إجابة إيجابية (أي أن Ui=1) للبند كوظيفة "θ" ويرمز إليها اعتيادياً بـ "P (θ)". في هذا الفصل نركز على النموذج التوريدي ثنائي الأبعاد الذي يستخدم الشكل الرياضي.

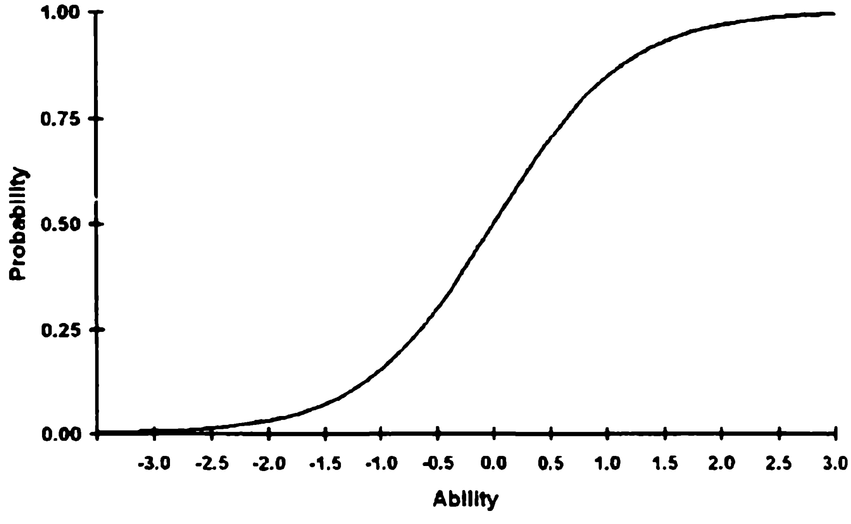
$$P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)}$$

في المعادلة السابقة تكون D مجموعة ثابتة تعادل 1.702 لأسباب تاريخية (أي بحيث أن يطابق النموذج التوريدي بدقة وظيفة إجابة البند logistic الطبيعية) =Ex (ExP) [X] حيث الثابت الرياضي يساوي تقريباً 2.718 في هذه المعادلة تكون ai وbi نواظم البند التي تكون ذات أهمية جوهرية بالنسبة لنا. ومؤشر التباين (التمييز) للبند ويرمز إليه بـ (ai) يدل على حدة ميل إجابة البند، وإجابة الميل التي

ترتفع بزاوية حادة الميل في فترة ما تمكننا من التمييز بدقة بين المجيبين الذين تكون لديهم "θs" أدنى أو المجيبين الذين تكون لديهم "θs" أعلى بعض الشيء. ومؤشر صعوبة البند bi مؤشر مكاني حيث θ=bi يجب الأخذ بعين الاعتبار أن Pi(θ) = 0.5 بحيث تكون bi النقطة على الاستمرارية الخفية الكامنة حيث يكون للمجيبين فرصة 50٪ للإجابة.

يعرض الشكل 1.11 وظيفة الإجابة للبند بالنسبة لبند افتراضي في مستويات q الدنيا تدل وظيفة إجابة البند على أن المجيبين سيكون لديهم احتمال تبني البند يقرب من الصفر، ويكون للمجيبين احتمالات إجابة إيجابية تقترب الواحد لدى قيم q العالية وبالتالي يقدم البند الافتراضي تمييزاً جيداً بين المجيبين الذين تكون لديهم قيمة q منخفضة بشكل معتدل ومرتفعة بشكل معتدل.

يسلك مسار استخدام نظرية إجابة البند لتقييم تعادل القياس لبند ما بين مجموعتين ثقافتين الاتجاه التالي: أولاً يجب جمع عينات ممثلة لذلك من كل من أفراد كل من الثقافتين، ثم نقدر نواظم البند بشكل مستقل للمجموعتين.



الشكل رقم 11-1

منحنى مميز لبند مفترض

وبعد وصل قياسات θ للمجموعتين (أي وضع تقديرات للنواظم على المقياس ذاته) يمكننا أن نقارن ما إذا كانت إجابة البند للثقافة الأولى تختلف بشكل ذي دلالة عن وظيفة إجابة البند للثقافة الثانية. يتم هذا التقييم باختبار أهمية متعدد التباينات يقارن تقديرات ai وb من عينة المستجيبين من الثقافة الأولى مع ما تم تقديره لـ ai وb من الثقافة الثانية. في الفصل التالي تم وصف كل خطوة من التحليل الذي استخدمناه لدراسة التعادل بين الثقافات في المقاييس المكيفة بتفصيل أكبر.

خطوات في تحليل المعلومات:

البُعدية: في الجزء الأكبر استخدمت الدراسات التي تبحث في تعادل القياسات عبر الثقافات عن طريق نظرية إجابة البند، نماذج وحيدة (أحادية) الأبعاد (أي نماذج تظهر فيها - ذات تدريجات وليس "vector" في البحث الجاري وصفة في هذا الفصل نستخدم نموذج "logistic" ذات ناظمين يكون أحادي الأبعاد وبذلك من الضروري اختبار الدرجة التي تنطبق فيها المعلومات على الافتراض أن هناك طرقاً عديدة لدراسة البعدية "Dimensionality". والظاهر أن هناك طرقاً عديدة لتقييم البُعدية بعدد الأشخاص الذين يقومون بعمل القياس النفسي. قدم "هاتي 1985" سرداً مفصلاً للطرق التي يمكن باستخدامها قياس البُعدية.

والطريقة التي نستخدمها عادة لقياس البعدية طريقة بسيطة؛ تحسب التناسقات ثلاثية الوظائف tetrachoric من أجل البند المدرج بشكل ثنائي وبعد ذلك تحلل المصفوفة(Matrix) الارتباط بالتحليل العاملي للمحور الرئيس، وبشرط أنه لا يوجد من العوامل ما هو متطرف أكثر مما يجب (أي أن كافة المستجيبين تقريباً يجيبون سلبياً، أو كل المستجيبين يجيبون إيجابياً) وتقدم أول قيمة "eigen" التي تكون عالية نسبة لقيمة "eigen" الثانية تقدم شاهداً جيداً على عاملٍ كامنٍ مهيمن. كما أن هناك دعماً آخر يأتينا للبُعدية الأحادية إذا كانت كل البنود لها ثقل عالي القيمة على العامل الأول (غير المدور).

تجدر الملاحظة أنه ليس من الضروري وجود مجموعة من البنود تكون أحادية البعد بشكل مطلق من أجل التطبيقات العملية لنظرية الإجابة للبند، وإنما بيّن كل من دراسكو وبارسونز 1983 وريشكاسه 1979 وينكر وستاوت 1994 أن عنصراً مهيمناً وحيداً كاف.

تقدير متغيرات البند (Item Parameter Estimation)

يفترض أن تحليل العوامل يعطي عاملاً مهيمناً فريداً نقدر بعد ذلك نواظم البنود. بينت دراسات التشابه مثل (دراسكو 1989 وماكلاخلان ودراسكو 1987) أن التقييم الهامشي لـ "بوك" (بوك وإيكن 1981 وبوك وليبرمان 1970) يجب أن يستخدم هنا. يقدم لنا برنامج الكمبيوتر Bilog (ميزليفي وبوك 1989) خيارين من أجل التقييم الهامشي: الاحتمالية المطلقة (العظمى) واحتمالية Bayesean"بيزيان". تقدم طريقتا التقييم المذكورتان، باستخدام عينات كبيرة، نتائج متشابهة، ولكن مع العينات الصغيرة تكون تقييمات بيزيان أكثر ثباتاً. عادة نقيم متغيرات البند في الحالتين ونأخذ عينات ذات حجم متواضع نسبياً (200-300).

كثيراً ما نشاهد أن واحداً أو أكثر من متغيرات البند المقيمة باحتمالية عظمى تكون متطرفة أكثر مما يمكن تصديقها (مثلاً تباين مقدر يبلغ 3.0) فمثل هذه التقييمات المتطرفة يكون لها عادة أخطاء معيارية عالية جداً.

أنسب طريقة لتفسير متغير تباين الاحتمالية العظمى يبلغ 3.0 ذو خطأ معياري يبلغ 1.2، هو أن عينة أكبر ضرورية من أجل تقييم الاحتمالية العظمى. ولنفترض وجود صعوبة، كالعودة إلى الهند مثلاً لجمع المزيد من المعلومات، نعود إلى تقييم بيزيان تكون تقييمات بيزين عادة أقل تطرفاً، فناظم قدر بـ 3.0 بالاحتمال الأعظم قد يقيّم بـ 1.3 وتكون له أخطاء معيارية أقل. والخطأ المعياري للاحتمالية العظمى الذي يبلغ 1.2 قد يكون 0.20 بهذا التقييم (بيزيان).

الرسوم البيانية الملائمة (Fit Plots)

بعد دراسة نواظم البنود من المهم اختبار المدى الذي يصف فيه نموذجنا من نظرية الإجابة للبند بشكل وافٍ إجابات البند. هناك عدد من الطرق المتوفرة من أجل هذا الهدف. وقد وجدنا أحدها واسمه "الرسوم البيانية الملائمة" ذات فائدة بشكل خاص. فبشكل عام تقارن الرسوم البيانية الملائمة النسبة الفعلية من إجابة المستجيبين لبند ما في فترة θ ثيتا حتى نسبة مقدرة، (أي إلى وظيفة الإجابة للبند) وصعوبة بناء حبكة مناسبة هي أننا لا نعرف قيمة θ ثيتا لكل مجيب. كما أنه من أجل مقياس قصير نتوقع تفاوتات كبيرة نسبياً بين الـ θ ثيتا المقدرة وتلك التي تتبع لكل مجيب.

وصف "دراسكو" و"ليفين" و"تساين" و"وليامز" و"ميد" (1995) أسلوباً يتصدى للمشكلة، فبدلاً من توزيع كل مجيب على زمرة θ وحيدة مبنية على تقرير قابل للنقض، يوزع كل مجيب بشكل متناسب على فئات θ المتعددة بشكل مبني على الاحتمالية الخلفية التي تقع فيها الـ θ ثيتا التابعة لذلك المجيب في تلك المدة و"العدّ غير الحقيقي" لعدد المجيبين الذين يصادف أن يكونوا في تلك المدة للحصول على "نسب إجرائية" تتبنى البنود، هذه النسبة الإجرائية تقارن عندها بوظيفة الإجابة للبند IRF ، انظر دراسكو وصحبه من أجل التفاصيل الفنية لهذا الإجراء، وانظر الأوراق التي أصدرها "ستون" (ستون 2003- ستون: 2000) من أجل تفاصيل الإجراء المتعلق بها.

مصفوفات التوصيل(linking Metrics):

يستنتج برنامج الكمبيوتر Bilob الافتراض المستنتج أن الخاصية الكامنة لهذا التوزيع الطبيعي المعياري في مجموع السكان الذين أخذت العينة منهم، إلا أنه لا يوجد سبب ملزم مسبقاً يجعلنا نعتقد أن المجموعات السكانية لديها الانحراف المعياري. (أي واحد)، والمعدل ذاته (أي الصفر) لأي خاصية كامنة. ودون فقدان العمومية يمكن أن أو يقاس مجتمع ما بحيث يكون لديه المعدل صفر ووحدة

الانحراف المعياري، ولكن تدرج أو مقياس الخاصية الكامنة لمجتمع آخر يجب أن يوصل بمقياس المجتمع الأول.

اقترحت عدة طرق لوصل مصفوفات الخاصية الكامنة. "قدم لنا (سيغال 1983) عرضاً جيداً" وقد بين البحث (مثلاً سيغال 1983) أن الخط البياني لخواص الاختبار الذي قدمه ستوكنغ ولورد (1983) والذي يصل الأعمال بشكل جيد وبرنامج Equate (الذي قدمه "بيكر" و"القرني" و"الدوسري 1991) يقدمان أداة يمكن استخدامها لأداء هذا النوع من الوصل. يقدم برنامج بيكر وصحبه معاملات انحناء وتقاطع المعاملات (أ و ب على التوالي) لتحول خطي يصل المقاسات. والمعاملات الموصولة هي:

$$\hat{a}^* = \frac{\hat{a}}{A} \text{ and } \hat{b}^* = A \times \hat{b} + B.$$

تجدر الملاحظة أن عناصر التباين والتباين المشترك للمصفوفة المرافقة Matrics التغير والتغير المرافق لمؤشرات البند المقدرة يجب أن تحول أيضاً بشكل جيد.

$$Var(\hat{a}^*) = \frac{Var(\hat{a})}{A^2}, Var\left(\hat{b}^*\right) = A^2 \times Var\left(\hat{b}\right),$$
$$\text{and } Cov\left(\hat{a}^*, \hat{b}^*\right) = Cov\left(\hat{a}, \hat{b}\right).$$

ويصبح الوصل عملية تلقائية عندما لا تبدي أية بنود أداءً تفاضلياً عبر الثقافات، إلا أن بعض الأداء التفاضلي للبنود (DTF) يمكن أن يحدث حتى مع أفضل أشكال التكيفات، ويمكن أن يتشوش الوصل عندما تشمل البنود ذات الأداء التفاضلي في التحليل. وبالتالي يمكن استخدام الإجراء التكراري التالي. بعد وصل مبدئي ياختبار الأداء التفاضلي للبنود (DTF) من أجل كل بند (توصل الإجراءات لاحقاً)، ثم توضع البنود ذات إحصاءات الأداء التفاضلي للبنود (DTF) جانباً مؤقتاً ويعاد وصل القياسات باستخدام البنود التي ليس لديها أداء تفاضلي للبنود (DTF)

وبعد الوصل تعاد كافة إحصائيات الأداء التفاضلي للبنود ويستمر هذا الإجراء من وصل البنود التي ليس لها أداء تفاضلي DTF وإعادة إحصاء الأداء التفاضلي لكافة البنود إلى أن يتبين للمجموعة ذاتها من البنود حصول أداءٍ تفاضلي (للبنود) DTF في تكرارين متتالين. وقد بينت دراسات التماثل (التي أجراها مثلاً كاندل ودراسكو 1988) فعالية الوصل التكراري.

تحليلات أداء البنود التفاضلية متعددة المجموعات:

نظر لورد (1980) وغيره إلى الأداء التفاضلي للبنود Differential ITem Functioning (DIF) من منظور مقارنة مجموعة جوهرية بمجموعة مرجعية. إلا أنه عند توفر المعلومات في عدة مجموعات من أجل التحليل كان أداء اختبارات الدلالة من أجل الاندماجات الزوجية كافة يبدو متوازياً مع استخدام الاختبارات المتعددة من أجل مقارنة معدل المجموعة وليس من أجل تحليل عام للتباين يختبر في الوقت ذاته مساواة المعدلات لكل المجموعات. قام كيم وكوهن وباركن (1995) بمساهمة مهمة لدراسة الأداء التفاضلي للبند DIF، عندما قدموا طريقتهم في تحليل DIF عن طريق مربع Square بين مجموعتين (لورد 1977–1980)، وبشكل محدد، فإن الفرضية الصفرية من أجل اختبار المساواة بين النواظم للبند (i) فيما بين المجموعات هو

H_0: $C\xi_i = 0$, حيث Matrix C تضادي يحتوي على عدد P من صفوف:

$$\xi = (a_{i1}, b_{i1}, \ldots a_{iK}, b_{iK})',$$

هو Vector نواظم البنود، وإحصاءات الاختبار التي وضعها كيم وزملاؤه هي:

$$Q_i = (Cv_i)'(C\Sigma_i C')^{-1}(Cv_i),$$

حيث إن Vi هو vector يحتوي على تقديرات ناظم البند

$$v_i = \left(\hat{a}_{i1}, \hat{b}_{i1}, \ldots, \hat{a}_{iK}, \hat{b}_{iK}\right),$$

وP هو رتبة C الذي يكون عادة 2 (K - 1) والتوزيع غير المتلاقي Q هو مربع Chi-square بدرجات P من الحرية (كيم وصحبه 1995).

تحليل أداء الاختبار التفاضلي (Differential Test Functioning Analyses)

بالرغم من أنه من النافع جداً تقييم مدى الأداء التفاضلي على مستوى البند، ففي معظم التطبيقات يستخدم الباحثون علامة مقياس كلية وليس بنوداً مستقلة، وعلى ذلك فالأداء التفاضلي للبند DIF كـ DTF بحد ذاته قد لا يكون ذا أهمية في العديد من الظروف وإنما (DTF) أي أداء الاختبار التفاضلي هو الموضوع ذو الأهمية الرئيسة: أي هل يكون لمجيبين لديهم قيمة متساوية في ثيتا ولكن عيناتهم أخذت من ثقافات ويجيبون بلغات مختلفة - درجات كلية متساوية على مقياس ما؟

قدم "راجيو" و"فان درليندن" و"فلير" (1995) طريقة لاختبار (DTF) تحسب حساب آثار البند التعويضية مثل (DTF) الذي يعمل في اتجاه واحد يمكن إلغاؤه بفعل (DIF) على بند آخر يعمل في الاتجاه المعاكس وقيمه طريقة "راجيو" وصحبه في معرض ترجمة مقاييس تستخدم في بحث ما بين الثقافات تكمن في أن البنود التي تبدي (DIF) لا تحتاج حتمياً لأن تزال من المقياس وإنما لا تحتاج البنود للحذف إلا عندما يكون المؤشر (DTF) عالياً وذا دلالة إحصائية. من منظور تحليل (DTF) يمكن لإجراءات (DIF) إزالة بنود ذات قيمة دون حاجة لذلك، فليس هناك حاجة لحذف بنود عندما يكون (DTF) الكلي غير ظاهر.

يمكن استخدام برنامج DFITDUA الذي وضعه "راجيو" وصحبه (1995) لتقييم (DTF) بين أزواج من العينات (كلمة "راجيو" التي ألقاها بذاته في نيسان 2000) يقوم راجيو حالياً بوضع ملحق متعدد المجموعات لتحليل (DTF) وهو يعطي إحصاء كاي مربع Chi Square يستخدم لتقييم أهمية (DTF) بين مجموعتين. عندما يكون مؤشر (DTF) ذا دلالة إحصائية يتعرف برنامج DFITDUA على البنود ذات المساهمة الأعلى لمربع كاي Chi Square ويزيلها.

وجد فلير (1993) أن مؤشر (DTF) كان حساساً بشكل زائد لأحجام العينات الكبرى.

مع هذه الحساسية يوصي راجيو وصحبه (1995) بعدم حذف البنود التي ينتج عنها مربع كاي Chi إلا عندما يكون مؤشر (DTF) الكلي أعلى من 0,006.

ملخص: يظهر أن الإجراء التحليلي الموصوف أعلاه يقدم وسيلة ذات نفع في تقييم التساوي في القياس بالمقاييس المكيفة للاستخدام بين اللغات والثقافات المتعددة، فهو يقيس افتراض بعدية IRT ويدرس مناسبة النموذج المقدر. ثم استخدمت الطرق المدخلة حديثاً التي تقارن تقييمات البنود بين عدة مجموعات وDTF كلية. لتقديم توضيح لهذا الإجراء نقدم فيما يلي اختباراً لتعادل القياسات في أربعة من المقاييس وضعت أصلاً في الولايات المتحدة وترجمت فيما بعد إلى اللغات الإسبانية والبولونية والمراثي ليتم استخدامها في المكسيك وبولونيا والهند على التوالي.

المنهج METHOD

العينات

كجزء من دراسة أكبر تبحث في فعالية الإجراءات المستخدمة بشأن الموارد البشرية فيما بين الثقافات، طبقت دراسة استطلاعية على 939 من الموظفين والعاملين في شركة طباعة ونشر متعددة الجنسيات، طبقت الدراسات على الموظفين في مؤسسات مقرها الولايات المتحدة (عدد أفراد العينة 239 = n)، والمكسيك (عدد أفراد العينة 253 = n) والهند (عدد أفراد العينة 201 = n) وبولندة (عدد أفراد العينة 246 = n) كانت الغالبية العظمى للمجيبين من مستخدمي اللغة الواحدة. وقد أخذت الغالبية العظمى في عينات المجيبين من الإدارة والعاملين في الإنتاج ومختلف المستويات الإدارية حسب التسلسل الهرمي. كما أبلغ المجيبون بأن إجاباتهم ستكون سرية بشكل مطلق وأن مساهمتهم تطوعية بحتة بإرادتهم فقط. كما أنهم أعلموا بأن إدارات مؤسساتهم والمستويات العليا من الإدارة في كل مؤسسة قد أعطت إذناً بإجراء هذه الدراسة ضمن أوقات العمل.

المقاييس:

قيم الاستبيان خواص المجيبين الديمغرافية (السكانية) وبيئتهم الأساسية (الخلفية)، كما أن البحث اشتمل على مقاييس ذات تدريجات لتقيس أنماط السلوك القيادية الموحية بالقوة وفرص التقدم المستمر والاستقلالية (العزلة) التنظيمية والالتزام التنظيمي.

وقد كانت المقاييس المستخدمة لقياس الجوانب المختلفة من الرضى عن العمل والضغط العام للوظيفة من الأمور التي أوليناها الاهتمام الخاص في هذا الفصل. وقد استخدمت نسخ ذات تسعة بنود من المقاييس الفرعية: الرضى عن العمل ذاته، والرضى بالإشراف والرضى بالعاملين مع الشخص المجيب، وهي مأخوذة من دليل توصيف العمل (JDI) بقلم ("سميث" و"كيندال" و"هونين" 1969) منقحة من قبل ("رزونوفسكي" 1989). وقد استعمل المشاركون المجيبون مقياس إجابة مدرج بثلاث درجات من "نعم" أو "لا" أو "لا أعرف" ليبينوا مدى وصف أي من الصفات أو العبارات لوظيفتهم. يحتوي المقياس العام (الذي وضعه "سيدمان" و"ماكراري" عام 1992) على صيغة إجابة تماثل صيغة (JDI) ويقيس التوتر الوظيفي العام. يحتوي الجدول 11-1 على المقاييس والبنود التابعة لها بأكملها.

التحليلات

تسجيل المتغيرات:

سجلت الإجابات لمقياس (JDI) في علامات ثنائية كما هو مطلوب من أجل نموذج (IRT) المستخدم في هذه الدراسة. وقد أ عطيت درجة واحدة للبنود الإيجابية المجاب عليها بالإيجاب والبنود السلبية المجاب عليها بالنفي. أما الإيجاب بالنسبة للبنود السلبية وعدم الإيجاب مع البنود الإيجابية فقد كان الرمز المعطى لها صفراً. أما الإجابات المقابلة لإشارة الاستفهام فقد أعطيت علامة الصفر نتيجة لما وجد إجرائياً في أن مثل هذه الإجابات كانت مرتبطة بشدة بعدم الرضى بالوظيفة

أكثر من ارتباطها بالرضى بالوظيفة. (هانيش 1992- سميث 1969) وقد وضعت الدرجات بالنسبة لمقاييس "SIG" بشكل موازٍ، أما الإجابات التي تشير إلى توتر أعلى فكانت تعطى علامة /1/، والإجابات التي تبيّن قدراً أقل من التوتر تعطى علامة صفر. إلا أن إجابات إشارة الاستفهام فقد أعطيت علامة /1/. كما أن مؤشرات إجابات المقياس كافة التي تحتوي على نقص لأكثر من بند واحد قد استبعدت من التحليلات التالية فيما بعد.

الجدول رقم 11-1

بنود الرضى بالمشاركين في العمل وبالمشرف عليه وبالعمل ذاته من فهرس توصيف وظيفي ومقياس التوتر بشكل عام

	United States				Poland				Mexico				India			
	p	r_b	*a*	*b*	*p*	r_b	*a*	*b*	*p*	r_b	*a*	*b*	*p*	r_b	*a*	*b*
Coworker Satisfaction																
Boring	.840	.624	.922	–1.448	.854	.878	.241	–1.350	.923	.694	.907	–2.112	.796	.646	.948	–1.212
Slow	.690	.773	1.179	–.638	.758	.643	.890	–1.102	.774	.788	1.161	–.951	.773	.929	1.366	–.937
Loyal	.426	.636	1.084	.287	.536	.679	1.118	–.101	.557	.678	1.219	–.128	.617	.661	.969	–.409
Responsible	.706	.739	1.134	–.702	.686	.728	1.087	–.647	.812	.967	1.396	–1.046	.759	.819	1.156	–.927
Waste of time	.651	.752	1.170	–.498	.808	.794	1.108	–1.214	.484	.258	.593	.118	.630	.426	.729	–.545
Lazy	.634	.792	1.287	–.420	.741	.828	1.189	–.887	.759	.752	1.121	–.916	.790	1.005	1.465	–1.000
Unpleasant	.758	.648	.962	–.976	.845	.809	1.132	–1.422	.844	.786	1.058	–1.348	.693	.784	1.135	–.650
Intelligent	.664	.686	1.017	–.563	.683	.811	1.264	–.605	.747	.662	.971	–.898	.716	.336	1.031	–.774
Work well together	.664	.686	1.024	–.567	.716	.782	1.178	–.759	.780	.855	1.241	–.963	.819	.761	1.066	–1.241

(continued on next page)

TABLE 11.1 (continued)

	United States				Poland				Mexico				India			
	p	r_b	a	b	p	r_b	a	b	p	r_b	a	b	p	r_b	a	b
Supervisor Satisfaction																
Hard to please	.515	.770	1.198	–.027	.379	.539	.865	.525	.648	.726	1.171	–.446	.286	.304	.689	.958
Impolite	.690	.782	1.178	–.660	.780	.881	1.308	–1.302	.855	.992	1.413	–1.289	.534	.793	1.356	–.140
Praises good work	.414	.638	.969	.338	.610	.577	.821	–.431	.728	.514	.806	–.930	.609	.727	1.122	–.358
Tactful	.450	.479	.729	.231	.599	.748	1.080	–.317	.636	.519	.817	–.505	.441	.061	.508	.239
Annoying	.573	.816	1.289	–.222	.640	.873	1.351	–.484	.800	.915	1.257	–1.052	.537	.800	1.377	–.136
Bad	.715	.875	1.426	–.713	.751	.890	1.320	–.936	.861	.900	1.292	–1.311	.637	.771	1.291	–.491
Interferes with my work	.745	.775	1.176	–.882	.785	.803	1.141	–1.109	.588	.291	.592	–.372	.525	.775	1.233	–.107
Gives confusing directions	.603	.668	.965	–.365	.662	.863	1.294	–.552	.714	.761	1.164	–.676	.576	.772	1.202	–.268
Knows how to supervise	.477	.813	1.294	.100	.580	.934	1.575	–.247	.750	.754	1.168	–.851	.551	.656	.944	–.187

Work Satisfaction																
Fascinating	.303	.759	1.288	.721	.203	.762	1.226	1.079	.604	.569	.915	−.353	.553	.511	.832	−.234
Satisfying	.601	.865	1.343	−.340	.543	.709	1.119	−.167	.783	.789	1.209	−.989	.822	.952	1.401	−1.183
Boring	.657	.729	.999	−.597	.648	.784	1.328	−.512	.849	.667	.982	−1.437	.777	.713	1.001	−1.152
Creative	.422	.773	1.151	.294	.244	.660	1.032	.941	.744	.600	.954	−.879	.559	.599	.942	−.262
Challenging	.686	.719	.996	−.721	.823	.413	.758	−1.564	.550	.311	.649	−.209	.697	.808	1.176	−.720
Gives sense of accomplish-ment	.644	.924	1.518	−.478	.404	.737	1.134	.319	.838	.814	1.219	−1.235	.731	.932	1.441	−.767
A source of pleasure	.343	.786	1.315	.570	.245	.733	1.179	.912	.500	.552	.999	.026	.620	.485	.789	−.506
Dull	.665	.846	1.233	−.586	.438	.642	.985	.221	.500	.589	1.000	.002	.620	.485	.789	−.506
Interesting	.628	.966	1.678	−.409	.534	.841	1.412	−.118	.846	.862	1.252	−1.271	.680	.730	1.058	−.723

(continued on next page)

TABLE 11.1 (continued)

	United States				Poland				Mexico				India			
	p	r_b	a	b	p	r_b	a	b	p	r_b	a	b	p	r_b	a	b
Stress in General																
Hectic	.612	.682	.975	–.432	.409	.527	.910	.282	.576	.452	.850	–.306	.589	.644	1.022	–.364
Tense	.565	.850	1.371	–.218	.641	.533	.894	–.566	.282	.696	1.131	.742	.481	.794	1.216	.032
Frantic	.278	.821	1.367	.788	.298	.460	.869	.701	.070	.534	.895	2.091	.350	.829	1.294	.507
Pressured	.648	.774	1.186	–.523	.813	.018	.530	–1.841	.372	.753	1.358	.381	.508	.841	1.313	–.051
Hassled	.333	.727	1.092	.616	.495	.572	.955	–.036	.080	.586	.917	2.000	.406	.882	1.409	.287
Relaxed	.638	.767	1.164	–.494	.840	.230	.680	–1.1661	.344	.628	1.138	.531	.520	.256	.574	–.154
Many things stressful	.612	.811	1.256	–.394	.583	.682	1.179	–.308	.448	.510	.858	.169	.500	.860	1.346	–.024
Nerve-wracking	.381	.845	1.390	.400	.473	.656	1.141	–.051	.162	.739	1.089	1.292	.307	.809	1.238	.649
More stressful than I'd like	.475	.744	1.075	.095	.378	.643	1.109	.327	.284	.304	.660	.979	.369	.778	1.048	.455

اختبار أحادية الأبعاد

من أجل اختبار مناسبة افتراض أحادية الأبعاد بالنسبة للمقايس الأربعة أجريت تحليلات للعوامل الأساسية (الرئيسة) تكرارياً وذلك في كل بلد كما وصفنا سابقاً. وقد استخلصت العوامل ذات قيمة eigen التي تكون أعلى من 1.0 وإذا حصلنا على استخلاصات وحيدة العامل أو شحنات عالية على عامل المحور الرئيس الأول (PAF) فذلك يعني أن مقياس "ai" يقيس تركيبة مهيمنة واحدة وذلك يدعم الافتراض بأحادية البعد.

تقييم متغيرات البنود Item Parameter Estimates

استخدم بيلوغ "Bilog" (مليفي وبوك 1989) لتقييم نواظم البنود للنموذج التنظيمي ذي الناظمين بالنسبة لكل مقياس في كل من العينات الأربع. حدد المدى الأقصى لدورات الـ(EM) المئة ولدورات "نيوتن -رافسون" العشر وقد استخدمت 30 نقطة تربيعية بدلاً عن العدد الناقص وحدد مبدأ الدمج ذو قيمة 0.001 وحدد توزيع مسبق طبيعي (0.1) N من أجل ناظم صعوبات البنود، بالإضافة إلى ذلك حدد توزيع مسبق طبيعي جذعي، (0,02) N من أجل نواظم التفريق بين البنود.

وصل القياسات تكرارياً

استخدم برنامج بيكرو (1991) "EQUATE" لوصل قياسات كل مجموعة مركزية (المكسيك، وبولندة والهند) إلى مجموعة المرجع (الولايات المتحدة) بطريقة ستوكينغ -لورد (1983)، كما استخدم الإجراء التكراري الذي وصفه كانديل ودراسكو (1988) وليونغ ودراسكو (1986) في إجراء الوصل، وهنا استخرجت مجموعة مبدئية لمعاملات الوصل لتحويل تقييمات النواظم من كل في المجموعات المركزية (موضوع الدراسة) إلى قياسات مجموعة المرجع. أجري تحليل (DIF) متعدد المجموعات لـ (كيم وصحبه 1995) واستبعدت البنود التي وجد أنها تبدي

(DIF) في مجموعة البنود المدروسة وطبقت طريقة "ستوكينغ -لورد" في الوصل ثانية باستخدام البنود غير المتأثرة بالانحراف فقط؛ وبعد إعادة الوصل أعيد اختبار كافة البنود من أجل (DIF) وتم تكرار هذا الأسلوب إلى أن ظهرت المجموعة ذاتها من البنود (DIF) في محاولتين متتاليتين أجري تحليل (DTF) بعد وصل تقييمات البنود في المجموعات الثلاث موضوع الدراسة بمجموعة المرجع بطريقة الإجراء التكراري الموصوف سابقاً. وهنا قورنت كل من المجموعات الثلاث موضوع الدراسة (المكسيك، بولندة، الهند) بالمجموعة المرجع (الولايات المتحدة) لتقرير ما إذا كان هناك وجود لأداء المقياس التفاضلي الكلي لكل من المقاييس الأربعة.

النتائج

أوردنا إحصائيات نظرية الاختبار التقليدية، أي نسبة "الصحيح" وارتباطات مجموع البنود ذات السلسلة الثنائية من أجل التوتر في المقياس العام ومن أجل مقاييس الرضى بالعمل والرضى بالمشرف والرضى بالعاملين المرافقين في الجدول رقم 11-1 .

البُعدية Dimensionality

بينت التحليلات العاملية المجراة في كل عينة من أجل كل مقياس بأن المقاييس كانت أحادية الأبعاد بما فيه الكفاية للشروع بتحليلات (IRT) وقد بينّا نتائج (PAF) في الجدول 11-2 .

الرضى بزملاء العمل المرافقين

نتج عن تحليل العاملين لمقياس الرضى بزملاء العمل استخراج عامل وحيد في كل عينة. وقد فسر هذا العامل أكثر من ثلث التباين في كل عينة. وقد كان حمل البنود عالياً بعض الشيء في الولايات المتحدة يتراوح بين /0.47/ (ممل) إلى /0.68/ (كسول) (متراخ)، وتباينت الأحمال في المكسيك بين /0.43/ (ممل) و/0.75/ (يحمل المسؤولية) وتراوحت الأحمال في "الهند" من /0.36/ (تضييع

للوقت) إلى /0.83/ (كسول) وأخيراً في بولندة كانت الأحمال أيضاً عالية تراوحت بين /0.51/ (بطيء) و/0.67/ (ذكي).

الرضى بالمشرف على العمل:

نتج التحليل العاملي لمقياس الرضى بالمشرف على العمل المكون من تسعة بنود عن استخلاص عامل وحيد أيضاً في كل عينة مما فسر تبايناً يزيد عن الثلث في كل عينة. وقد كانت حمولات كل بند عالية شيئاً ما تتراوح بين /0.41/ (لبق) إلى /0.73/ (سيئ) في الولايات المتحدة. وهي المكسيك بين /0.25/ (يتدخل في عملي) إلى /0.71/ (يعرف كيف يشرف على الناس) وفي الهند بين /0.08/ (لبق) إلى /0.78/ (مزعج) وفي بولندة /0.46/ (صعب إرضاؤه) إلى (يعرف كيف يشرف على الناس).

الرضى بالعمل ذاته:

أدى التحليل العاملي لمقياس الرضى بالعمل، المكون من تسعة بنود إلى استخلاص عامل وحيد في كل من العينات الأربع. وقد تراوح التباين بموجب هذا العامل بين 25.5٪ في المكسيك إلى 38.7٪ في الولايات المتحدة. وقد كانت الحمولات الكلية في كل عينة عالية بعض الشيء، فقد تراوحت بين /0.16/ (ممل) و/0.80/ (مشوق) (يدعو للاهتمام) في الولايات المتحدة. وفي المكسيك تفاوتت الأحمال بين /0.14/ (ممل) و/0.77/ (مشوق يدعو للاهتمام) وفي الهند تروحت الأحمال بين /0.31/ (ممل) و/0.76/ (مرضي)، وأخيراً في بولندة تراوحت الأحمال بين /0.06/ (ممل) و/0.72/ (مشوق). ومن الممتع ملاحظة الميل للتقارب بين الحمولات لعوامل بين العينات، ففي كل دولة كانت البنود (ممل) و(كئيب) حمولة منخفضة بشكل متناسق بينما كان البند (مشوق) دائماً حمولة عالية على العامل العام.

الضغط العام للعمل (التوتر):

نتج عن تحليل PAF لمقـيـاس "التـوتر بشكل عـام" المكون من تسـعـة بنود (تدريجات) أيضاً عامل وحيد استخلص من كل عينة، وقد تراوحت التباينات المقاسة من 25.7٪ في المكسيك إلى/0.45٪ في الهند وكانت الحمولات العاملية عالية في كل من العينات في الولايات المتحدة تراوحت الحمولات من 0.56 مضطرب إلى 0.79 (متوتر).

الجدول رقم 11-2

نتائج تحليلات للعاملية للمحور الرئيس من أجل الرضى بالمشاركين في العمل وبالمشرف وبالعمل ذاته ومعدلات التوتر بشكل عام بحسب الدولة.

	United States		Poland		Mexico		India	
	# factors	% variance	# factors	% variance	# factors	% variance	# factors	% variance
:ale								
oworker	1	35.9	1	37.9	1	35.2	1	37.9
ipervisor	1	40.0	1	42.8	1	33.8	1	37.6
'ork	1	38.7	1	29.4	1	25.5	1	33.0
›b Stress	1	42.8	1	26.3	1	25.7	1	45.0

وتراوحت الحمولات في المكسيك بين /0.28/ (مضطرب) و/0.61/ (محطم للأعصاب) وفي الهند تراوحت الحمولات بين /1.19/ (مريح) و/0.77/ (متوتر) وأخيراً في بولندة كانت الحمولات تتراوح بين /0.13/ (مريح) و/0.63/ (محطم للأعصاب) فهنا أيضاً كانت هناك درجة من الاتساق بين العينات فيما يتعلق بالحمولات العالية والحمولات المنخفضة للبنود.

تقييمات ناظم البند:

وضعت الإجابات المعطاة علامات ثنائية والمستخلصة في 939 موظفاً كما وصفنا سابقاً كمدخل في Bilog وكان النموذج الدلالي قيد قيّم من أجل كل مقياس

باستخدام المعلومات المأخوذة من كل قطر. وقد وضعت تقييمات النواظم الناتجة في الجدول 11-1 وجرت مناقشتها فيما يلي:

الرضى بزملاء العمل

بينت القياسات المبدئية (أي قبل التعادل) Bilog للشعور بالرضى عن زملاء العمل في كل من الدول الأربع أن البنود كانت بشكل عام (سهلة) (أي أدخلت إيجابياً في الكثير من المرات) وكانت كافة التقييمات (فيما عدا اثنين) ويبلغ تعدادها 36 ذات تقييمات سلبية. وهكذا فالأفراد ذوو مستويات الرضى المتراوحة بين نصف وواحد انحراف معياري (SD) تحت المعدل كان لديهم احتمال بنسبة 50٪ الادخال العديد من البنود إيجابياً (أي تبيان الرضى). كان معدل تقييم في الولايات المتحدة -614.- (SD = 0.416) وكان يتراوح بين (1.448 - ممل) إلى (0.287 مُوَالٍ) وكان معدل تقييم الصعوبة في المكسيك 916.- (SD = 0.645). وفي الهند، كان معدل التقييم 0.844 (SD = 0.42). أما تقييم النواظم فقد تراوح بين 1.241-(يعمل جيداً مع الآخرين) إلى 409.- (مُوَالٍ). أخيراً في بولندة كان المعدل 899.- (SD = 0.422) بالإضافة لذلك بدا أن البنود تميز بشكل جيد.

وكان معدل تقييم التمييز لبنود المشاركين في العمل في الولايات المتحدة 1.086 (SD = 0.117) ويتراوح بين 0.922 (ممل) إلى 1.287 (كسول). وكان المعدل في المكسيك لتقييم ai هو 1.074 (SD = 0.233)، وتراوحت التقييمات بين 0.593 (تضييع للوقت) إلى 1.396 (شعور بالمسؤولية) وكان معدل التقييم في الهند 1.096 (SD = 0.222) وكانت التقييمات تتراوح بين 0.729 (هدر للوقت) إلى 1.465 (كسول) أخيراً في بولندة كان معدل تقييم ai 1.134 (SD = 0.109) ويتراوح بين 0.890 (بطيء) إلى 1.264 (ذكي)، وهذه المعدلات تبين أن البنود تتمايز بشكل جيد.

الرضى بالمشرف على العمل:

كانت بنود الرضى بالمشرف بشكل عام (سهلة) أيضاً بالرغم من أن معدل تقييمات bi كان أكبر منه بالنسبة لبند الرضى بزملاء العمل. كان معدل صعوبة

البنود في الولايات المتحدة 244.- (SD= 0.439) وتراوحت التقديرات بين 882.- (يتداخل بعملي) إلى 0.338 (يمتدح العمل الجيد) وقد كان معدل تقييم b_i في بولندة 0.509 (SD= 0.500) وتراوحت تقييمات الناظم بين 1.302- (غير مهذب) إلى 0.525 (من الصعب إرضاؤه)، وفي المكسيك كان معدل الصعوبة 0.826- (SD= 0.351) وتراوحت التقييمات بين 1.311- (سيئ) إلى 0.372- (يتدخل في عملي)، أخيراً في الهند كان معدل لتقييم 0.054- (SD= 0.429) وتراوحت تقييمات الناظم بين 0.491- (سيئ) و0.958 (من الصعب إرضاؤه).

وباختبار تقييمات a_i نجد أن بنود الرضى بالمشرف تميز جيداً باستثناءات قليلة جداً في الولايات المتحدة كان تقييم معدل التمييز 1.136 (SD= 0.213) وتتراوح التقييمات بين 0.729 (لبق) و1.426 (سيئ). في بولندة كان معدل تقييم a_i 1.195 (SD=0.243) وكانت التقييمات تتراوح بين 0.821 (يمتدح العمل الجيد) و1.575 (يعرف كي يشرف على الناس) وكان معدل التمييز بين البنود في المكسيك 1.075 (SD= 0.272) وتراوحت التقييمات بين 0.592 (يتداخل بعملي) و1.413 (غير مهذب) أخيراً كان معدل تقييم a_i في الهند 1.080 (SD= 0.306) وتراوحت التقييمات بين 0.508 (لبق) و1.377 (مزعج).

الرضى بالعمل:

تظهر دراسة تقييمات ناظم صعوبة البنود أن هناك معدلاً واسعاً من صعوبة البنود في مجال قياس الرضى بالعمل وبالذات في الولايات المتحدة وبولندة، وتراوحت تقييمات صعوبة البنود في الهند والمكسيك على العموم في نهاية المجال المرقم بـ (سهل).

كان معدل صعوبة البند في الولايات المتحدة 172.- (SD = 0.547) وتراوحت التقييمات بين 721.- (يمثل تحدياً) و0.721 (ذو سحر) وكان معدل تقييمات b_i في بولندة 0.123 (SD= 0.840) وتترواح التقييمات بين 1.564- (يمثل تحدياً) و1.079

(ذو سحر) وفي المكسيك كان معدل تقييم bi .705- (SD = 0.576) حيث تتراوح التقييمات بين 1.437- (ممل) و0.026 (مصدر للسرور). أخيراً في الهند كان معدل صعوبة البند 697.- وتتراوح تقييمات ناظم البنود بين 1.183- (مرضي) و234.- (ذو سحر) في العينات الأربع. وقد كان معدل تمييز البند في الولايات المتحدة 1.280 (SD = 0.223) بتمييز للتقييمات يتراوح بين 0.996 (يمثل تحدياً) و1.678 (مشوق). وفي بولندة كان معدل تقييم ai 1.130 (SD = 0.194)، وكان تقييم ناظم البنود يتروواح بين 0.758 (يمثل تحدياً) و1.328 (ممل). وكان معدل تقييم التمييز في المكسيك 1.020 (SD = 0.189) حيث تتراوح تقييمات البند بين 649.- (يمثل تحدياً) و1.252 (مشوق). أخيراً في الهند كان معدل تقييم المنحى 1.113- (SD= 0.248) وتتراوح التقييمات بين 0.832 (ذو سحر) و1.441 (يعطي شعوراً بالإنجاز).

التوتر العام في الوظيفة:

أظهرت معايرة Bilog لمقياس التوتر بشكل عام المكون من 9 بنود انتشاراً جيداً لصعوبة البنود. كان معدل الصعوبة في الولايات المتحدة 0.018 (SD = 0.509) وتتروواح تقييمات ناظم البنود بين 523.- (عليه ضغط) و0.788 (يحطم الأعصاب). وكان معدل صعوبة البند 339.- (SD = 0.811) وكانت صعوبات البنود تتراوح بين 1.841- (عليه ضغط) و0.701 (يحطم الأعصاب). وفي المكسيك كان معدل تقييم bi 0.876 (SD =0.807) وتتروواح التقييمات بين 0.306- (مقلقل) إلى 2.091 (يدعو للجنون). أخيراً في الهند كان المعامل 0.149 (SD = 0.341) وتتروواح تقييمات الصعوبة بين 394.- (مقلقل) و0.649 (محطم للأعصاب).

وقد كانت تقييمات تمييز البنود بشكل عام جيدة جداً؛ فقد كان معدل تقييم ai في الولايات المتحدة 1.208 (SD = 0.148) والتقييمات تتراوح بين 0.975 (مقلقل) و1.390 (محطم للأعصاب). وفي بولندة كان معدل التقييم 0.919 (SD = 0.214)، وتتراوح التقييمات بين 0.530 (تحت ضغط) و1.179 (كثير من الأشياء متوترة)،

وفي المكسيك كان معدل تقييم ai 0.988 (SD = 0.209)، والتقيمات تتراوح بين 0.660 (أكثر توتراً مما أحب) و1.358 (تحت ضغط). أخيراً كان معدل التقييم في الهند 1.162 (SD = 0.256) والتقييمات تتراوح بين 0.574 (مريح) و1.409 (ملاحق).

وصل القياسات مع أداء البند التفاضلي متعدد المجموعات:

حيث إن تقييمات الناظم البنود الموصوف سابقاً عشوائية فيما يتعلق بمصدر الوحدة فقد وصلت قياسات كل عينة مركزية (موضوع الدراسة) للمجموعة بمجموعة المرجع باستخدام أسلوب الوصل التكراري الموصوف سابقاً. اختيرت الولايات المتحدة كمجموعة مرجعية لأن لغة المصدر لكل من المقاييس المستخدمة المكيفة هي اللغة الانكليزية. إلا أنه من المهم التنويه بأن أياً من هذه المجموعات موضوع الدراسة يمكن أن تختار كمجموعة مرجعية بالدرجة نفسها تقارب الوصل في تكرارين لكل عينة وكل مقياس. يمكن رؤية ثوابت التحويل في كل تحويل في الجدول 11-3. ونتبنى مقياس الرضى بالمشاركين في العمل لتوضيح إجراء الوصل التكراري. بعد الوصل المبدئي للمجموعات موضوع الدراسة بالمجموعة المرجعية (مثلاً A = 0.943 و B = 0.279 بالنسبة للمكسيك). طبق تحليل DIF متعدد المجموعات وجد DIF في بند واحد (هدر الوقت) لذلك استبعد هذا البند من المقياس وأعيد حساب معاملات الوصل باستخدام البنود التسعة الباقية. طبقت معاملات الوصل الجديدة (مثلاً: A = 1.037 و B = 0.446 بالنسبة للمكسيك) وأعيد حساب إحصاءات DIF فوجد أن بند (هدر الوقت) ما زال يبدي DIF وحيث أنه لا يوجد DIF بالنسبة لأي بند آخر انتهى إجراء الوصل التكراري.

الجدول 3.11

معاملات التحولات الخطية التي تصل المجموعات موضوع الدراسة بالمجموعة المرجعية من أجل كل بعد للوصف الوظيفي والتوتر الوظيفي

	Iteration 1		*Iteration 2*	
	A	*B*	*A*	*B*
Coworker Satisfaction				
Mexico	.943	.279	1.037	446
India	1.012	.272	1.081	.352
Poland	1.044	.334	1.059	291
Supervisor Satisfaction				
Mexico	.941	.540	1.015	.727
India	.932	-.170	.907	-.094
Poland	1.039	.288	1.046	.297
Work Satisfaction				
Mexico	.808	.422	.841	712
India	.905	.478	.902	.627
Poland	.851	-.305	.905	-.346
Job Stress				
Mexico	.779	-.687	.757	-.706
India	.971	-.172	1.062	-.167
Poland	.731	.176	.746	232

من المهم جداً تقدير قيمة كاي مربع Chi الحاسمة التي ستستخدم في تصنيف البنود على أنها تبدي DIF أو لا تبديها. يمكن لهذه القيمة الحاسمة أن تعتمد على عدد المجموعات المركزية (موضوع الدراسة) وعدد النواظم في نموذج IRT والنوع المرغوب فيه من مستوى Alpha وعدد البنود في المقياس. في هذه الدراسة كانت هناك ثلاث مجموعات موضوع الدراسة وناظمين وعلى ذلك فقد كان لإحصاء اختبار كاي مربع Chi ست درجات من الحرية إذا رغبنا في إبقاء Alpha كلية قيمتها 0.01 من أجل اختبار DIF على كل مقياس فإن تطبيق تصحيح بوفروني

سينتج ألفا Alpha لكل بند قيمتها 0.001 تقريباً ومربع كاي Chi حاسم قيمته 22.46 ومما تجدر ملاحظته أن كثيراً من البنود كان لها قيم مربع Chi تقرب من هذه القيمة (انظر الجدول 11-4 من أجل قيم مربع كاي Chi لكل بند من التكرارات 1 و2). وقد بينت حبكات IRF لمثل هذه البنود اختلافات صغيرة نسبياً، وعلى ذلك قررنا تصنيف البنود ذات / $\chi^2 \geq 50$ / على أنها قيم تبدي DIF تُظْهِر قِيمُ الـ (IRF) لمثل هذه البنود اختلافات عالية نسبياً). وبالنسبة لكل من المقاييس الثلاثة الأخرى (الرضى بالإشراف، الرضى بالعمل ذاته، التوتر بشكل عام) فإن عملية الوصل كررت مرتين فقط. وفي كل حالة وجد أن البنود المتعارف عليها على أنها منحازة في المرة الأولى كانت مماثلة للبنود التي وجد أنها منحازة في التكرار الثاني.

وكانت البنود التي أبدت (DIF) في المقاييس الأربعة قليلة نسبياً. وكان هناك بند واحد من مقياس "الرضى بالمشاركين في العمل" وجد أنه منحاز وهو "هدر الوقت" وكان له قيمة $\chi^2_{(6)}$ = 78.90 وهناك بند واحد من مقياس الرضى بالمشرف وُجد أنه يبدي () وهو: يتدخل في عملي وكان له $\chi^2_{(6)}$ = 93.23 وهناك بندان على مقياس الرضى قد أظهرا DIF وهما: «التحدي» وكانت له قيمة $\chi^2_{(6)}$ = 141.23 «الملل» وكانت له قيمة $\chi^2_{(6)}$ =086.48 . أخيراً وجد بندان من مقياس التوتر بشكل عام منحازين: (مقلِق) الذي كان له $\chi^2_{(6)}$ =111.43 و(مريح) بقيمة $\chi^2_{(6)}$ = 63.05.

ومن المهم أن نتذكر أن تحليل (DIF) متعدد المجموعات لا يقرر فقط أن هناك (DIF) بين مجموعتين أو أكثر من المجموعات تحت الدراسة. إن هذا التحليل لا يحدد أين ينشأ الأداء التفاضلي. ومن أجل تقرير ذلك يجب إجراء مقارنات ثنائية باستخدام مربع كاي Chi الذي قال به لورد.

تحليل أداء الاختبار التفاضلي:

من أجل تقييم الأداء التفاضلي على مستوى الاختبار، أجريت مقارنات ثنائية بين عينة الولايات المتحدة كمجموعة مرجعية، وكل من العيّنات الثلاث الأخرى موضوع الدرس. وبعد إجراء تقرير مستوى (DTF) بيّن برنامج DFITDUA البنود

التي يستحسن إزالتها للقضاء على الأداء التفاضلي الكلي للاختبار. وكان البند الذي يجري اختياره للاستبعاد هو الذي يساهم بشكل ذي دلالة في الـ DTF الكلي ويتم استبقاؤه إذا كان يؤدي إلى مؤشر DTF قيمته 0.006 أو أعلى (راجيو وصحبه 1995).

الجدول 11-4

احصاءات مربع كاي Chi للأداء التفاضلي للبند متعدد المجموعات للتكرارين 1 و2

	Iteration 1 χ^2	*Iteration 2* χ^2
Coworker Satisfaction		
Boring	11.75	10.42
Slow	6.09	5.64
Loyal	6.46	5.46
Responsible	10.36	5.98
Waste of Time	63.50*	78.90**
Lazy	4.41	4.79
Unpleasant	16.78	17.92
Intelligent	3.18	2.56
Work Well Together	8.03	5.02
Supervisor Satisfaction		
Hard to Please	39.30	35.71
Impolite	10.88	13.45
Praises Good Work	34.10	30.51
Tactful	15.95	13.43
Annoying	3.20	1.27
Bad	0.65	1.90
Interferes with my Work	76.04*	93.23**
Gives Confusing Directions	6.95	9.85
Knows How to Supervise	12.39	7.48
Work Satisfaction		
Fascinating	25.98	10.59
Satisfying	3.99	7.94

	Iteration 1 χ^2	Iteration 2 χ^2
Boring	7.40	10.33
Creative	31.02	12.41
Challenging	105.16*	141.23**
Gives Sense of Accomplishment	20.13	11.72
A Source of Pleasure	23.52	28.12
Dull	51.91*	86.48**
Interesting	16.53	13.84
Job Stress		
Hectic	51.53*	111.43**
Tense	4.16	2.57
Frantic	18.05	7.90
Pressured	16.91	22.58
Hassled	22.08	17.32
Relaxed	64.20*	63.05**
Many Things Stressful	28.84	23.68
Nerve-Wracking	1.12	3.64
More Stressful Than I'd Like	34.45	23.22

الرضى بالمشاركين في العمل:

لقد أظهرت المقارنة بين الولايات المتحدة والمكسيك أداتية اختبارية ذات دلالة: DTF = 0.034، $X^2(252) = 318.89$، $p < 0.01$. وقد تم اقتراح استبعاد أربعة بنود وهي «الملل»، «هدر الوقت»، «الكسل»، «العمل الجماعي الجيد». وعند استبعاد هذه البنود فإنه يمكن الحصول على قيمة دليل DTF = 0.004 و $X^2(252) = 705.81$، و $p < 0.001$ ، وتجدر الإشارة إلى أن قيمة مربع كاي Chi أصبحت أكبر وذات قيمة إحصائية واضحة؛ وفي الوقت نفسه فإن قيمة دليل DTF أصبحت أقل من 0.006 وطبقاً لبحث فلير (1993) فإنه يجب عدم استبعاد أي من البنود الأخرى.

وأظهرت المقارنات بين الولايات المتحدة والهند أدائية اختبارية تفاضلية ذات دلالة: DTF = 0.011، وX^2(200) = 958.09، و$p < 0.001$ وقد تم اقتراح استبعاد بند واحد وهو «هدرالوقت» وقد نتج عن استبعاد هذا البند مؤشر DTF = 0.002، وX^2(200) = 208.88.

وبالمقارنة بين الولايات المتحدة الأمريكية وبولندا ظهر أن هناك أدائية اختبارية تفاضلية ذات دلالة: DTF = 0.015، X^2(245) = 462.16، $p < 0.001$ وتم اقتراح استبعاد بند واحد وهو «هدر الوقت». وباستبعاد هذا البند فقد تم الحصول على مؤشر DTF = 0.000، وهكذا لم يتم استبعاد أية بنود أخرى.

الرضى بالإشراف على العمل:

بينت المقارنات بين الولايات المتحدة والمكسيك أدائية اختبارية تفاضلية ذات دلالة:

DTF =0.085 ، X^2 (252) = 1635.39، $p < 0.001$ وتم اقتراح استبعاد بند واحد وهو (يتدخل بعملي) ونتج عن استبعاد هذا البند مؤشر DTF قيمته 0.000.

وبينت المقارنات بين الولايات المتحدة والهند أدائية اختبارية تفاضلية ذات دلالة:

DTF = 0.055 ، X^2 (200) = 609.07، $p < 0.001$ وتم استبعاد بند واحد هو (يتدخل بعملي) ونتج عن حذف هذا البند مؤشر DTF قيمته 0.009 وX^2 (200) = 200.11 وعلى ذلك لم تحذف أية بنود أخرى.

ولم ينتج عن مقارنة الولايات المتحدة ببولندة أدائية اختبارية تفاضلية ذات دلالة، DTF = 0.002 وعلى ذلك لم تحذف أية بنود أخرى.

الرضى بالعمل نفسه:

بينت المقارنات بين الولايات المتحدة والمكسيك فروقات جوهرية فكانت القيم:

DTF = 0.466 , X^2(252) = 1568.25 ، $P < 0.001$ وتم اقتراح استبعاد بندين

هما: «التحدي» و«الملل». وباستبعاد هذين البندين تم التخلص من الفروقات (مؤشر DTF = 0.004).

وبينت المقارنات بين الولايات المتحدة والهند أيضاً اختلافات جسيمة فكانت القيم: DTF = 0.164، $X^2(200) = 518.94$، $P<0.001$. وتم اقتراح استبعاد بندين وهما: «التحدي و «الملل». ونتيجة لحذف هذين البندين تم الحصول على مؤشر DTF غير ذلك دلالة، $X^2(200) = 208.23$.

هذا ولم ينتج عن مقارنة الولايات المتحدة الأمريكية ببولندة أدائية اختبارية تفاضلية ذات دلالة، وكانت قيمة $X^2(245) = 272.86$.

التوتر بشكل عام:

بينت المقارنات بين الولايات المتحدة والمكسيك أدائية اختبارية تفاضلية ذات دلالة: DTF = 0.012, $X^2(252) = 560.61$, $P<0.001$ وتم اقتراح إزالة بند واحد هو (كثير من الأشياء شديدة التوتر) ونتج عن إزالة هذا البند مؤشر (DTF) قيمته 0.006 , $X^2(252) = 266.15$.

وبينت المقارنات بين الولايات المتحدة والهند عدم وجود أدائية اختبارية تفاضلية ذات دلالة وكانت قيمة $X^2(200) = 225.63$. وأخيراً نتجت عن المقارنات بين الولايات المتحدة وبولندة أدائية اختبارية تفاضلية ذات دلالة: $P<0.001$, 448.23 = DTF = 0.074, $X^2(245)$ وتم اقتراح استبعاد بندين وهما «مقلق» و«تحت الضغط» وعند استبعادهما من التحليل وجد مؤشر DTF غير ذي دلالة وأن قيمة $X^2(245) = 253.64$.

التوافق بين DIF للمجموعات المتعددة وتحليلات : DTF

قبل مناقشة هذه النتائج وما تتضمنه من أجل تكييف المقاييس، من الممتع دراسة التناسق (أي التوافق) العالي النسبة بين نوعي التحليل الأدائي التفاضلي. وبشكل خاص بينت التحليلات DTF وDIF متعددة المجموعات بدقة الدرجة ذاتها في بند الرضى بالإشراف (يتدخل في عملي) وانحياز بنود الرضى بالعمل ذاته (يمثل تحدياً) و(ممل) التي ظهرت فيها منحازة بالإضافة لذلك بين بند الرضى بالعمل ذاته (هدر الوقت) DIF في كلا تحليلي DTF وDIF للمجموعات المتعددة.

وكان الفرق الوحيد بين نوعي التحليل أن ثلاثة بنود إضافية اقترحت إزالتها عند المقارنة بين الولايات المتحدة والعينة المكسيكية. أخيراً كانت نتائج مقياس التوتر بشكل عام متناسقة بين طريقتي DIF وتم التعرف على بندي (مقلقل) و(مريح) على أنهما بندان يتصفان بـ DIF في تحليل DIF متعدد المجموعات. وقد وجد أن بند (مقلقل) كان يسهم في حصول DTF بين الولايات المتحدة وبولندة وكان بند (مريح) يسهم في حصول DTF بين الولايات المتحدة والهند. إلا أن تحليل DTF أيضاً أوحى بوجوب إزالة بندين آخرين عند مقارنة الولايات المتحدة بالمكسيك والولايات المتحدة ببولندة.

المناقشة:

قمنا في هذا الفصل بوصف توجه لتاختبار التعادل بين المقاييس المكيفة وأشكالها الأصلية إجرائياً. تجدر الملاحظة أن البدائل متوفرة لكل خطوة من هذا الإجراء التحليلي. فهناك طرق متعددة لدراسة البُعدية وتقييم نواظم البنود IRT واختبار مناسبة للنموذج المقيم ووصل القياسات ومعرفة كمية حجم DIF وقد اختيرت الأساليب في تحليلنا لتبيّن كفاءتها (جيدة العلم) في الدراسات المتماثلة (مثلاً التقييم الهامشي - الوصل التكراري) أو لأنها تختبر لنا فرضية معينة تهمنا (مثل كيم وصحبه 1995) DIF متعدد المجموعات: "راجيو وصحبه (1995) وDTF" ولكن تجدر الملاحظة أن الطرق الأخرى يمكن أن تعمل أيضاً أو أنها تَفْضُلُ هذه الطرق. كما أنه لم يسبق أن أجري بحث مُمُاثل يقارن بين الفعالية النسبية لمجموعات التحليلات المطلوبة من أجل اختبار تكيفات المقاييس والاختبارات. يمكن لدراسة تماثلية واسعة أن تتقاطع عاملياً مع طرق أخرى عديدة من أجل تقييم نواظم البنود وكذلك طرق عديدة من أجل اختبار مناسبة النموذج المقيم وهناك بدائل للوصل وطرقاً لتقييم DIF. يقدم التحليل العاملي طريقة بديلة لدراسة التكيفات. وعادياً تفترض نماذج التحليل العاملي أن المتغيرات الظاهرة ترتبط خطياً

بالبنى الكامنة وأن المتغيرات الظاهرة تتبع توزيعاً متعدد الاختلافات طبيعياً (اعتيادياً). ولكن هذه الافتراضات تنتهك بشكل مزرٍ بفعل البنود ذوات الدرجات الثنائية وعلى ذلك لا يتوقع الحصول على نتائج ذات معنى من التحليل العاملي. إلا أن بنداً ذا سبع فئات في الإجابة أو حتى خمس منها يمكن أن يقدم مثالاً تقريباً جيداً على التوزيع الطبيعي في بعض الظروف (أي عندما تكرس الفئات الوسطى ولا يكرس الفئات المتطرفة إلا القليل من الناس نسبياً -انظر دراسكو دورانز 1982) في مثل هذه الحالات يمكن لتحليل سوداريوم (1974)(MACS) الذي يتناول تحليل المعدل والبنية مشتركة التباينات أن يقدم بديلاً جيداً عن طرق IRT.

كلتا الطريقتين (IRT) المستخدمة هنا و(MACS) تقدمان للباحثين أداة نافعة لأنهما تقرران دقة تمثيل الأشكال المكيفة (في المقاييس) بالتعرف على البنود التي تقيس بشكل متعادل بين الثقافات المختلفة والثقافات التي لا تعطي مثل هذه القياسات. فبفرض أنه لا توجد بنود كثيرة تبدي (DIF) يمكننا مقارنة مفهوم الرضى بالعمل ذاته مثلاً بين الثقافات باستخدام البنود التي لا تبدي (DIF) فقط. ولكن بما أن الثقافات المقارنة معاً تتزايد باضطراد، يبدو أنه من المحتمل أن يصبح عدد البنود التي لا تبدي DIF قليلاً بحيث لا يمكن إجراء مقارنة ذات معنى باستخدامها. في مثل هذه الحالة كيف يتسنى للباحثين إجراء مقارنات ذات قيمة فعلية عن الرضى بالعمل بين الثقافات؟ نقترح استخدام إجراءات تعادل الاختبارات (انظر كولين وبرينان 1995) لوصل ما نلاحظه من قياسات الدرجات إحصائياً. في هذا التحليل يمكن للباحث استخدام تحليل DIF الذي قدمه راجيو وصحبه (1995) للتعرف على المجموعة الثانوية من البنود التي تقدم قياساً معادلاً من أجل كل ثقافة مدروسة مع الثقافة المرجع. يمكن أن تعتبر هذه البند (بنوداً مشتركة) في معادلة البنود. ويمكن للمجيبين في الثقافتين أن يُعتبروا (مجموعات غير متعادلة "متساوية") والمعلومات يمكن النظر إليها على أنها تم الحصول عليها من تصميم أخذ عينات ذي البند المشترك في مجموعة غير (متعادلة متساوية) (انظر ص 18-

21 كولين وبينان 1995)، ومن هنا فأي من إجراءات التعادل من أجل هذا التصميم يمكن أن يستخدم لإعادة تدريج الدرجات من أجل النسخة المكيفة للمقياس إلى قياس لغة المصدر. بعد مثل تلك المعادلة سيكون من الممكن مقارنة مثلاً درجات الرضى بالعمل، والاستنتاج بأن العاملين في ثقافة ما أكثر رضى من العاملين في ثقافة أخرى. وبالطبع سيكون المدى الذي تمثل فيه العينات المنتقاة من ثقافات المرجع والثقافات موضوع البحث التي تمثل المجتمعات التي تؤخذ منها موضوعاً جوهرياً، ولكن معادلة الاختبارات يمكن أن تقدم طريقة لحساب درجات يكون من الممكن مقارنتها بشكل فوري.

*****

المراجع

Bach, C. L. (1998, July). U.S. international transactions. *Survey of Current Business*, pp. 47–57.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15*, 78.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179–197.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77–90.

Drasgow, F., & Dorans, N. J. (1982). Robustness of estimators of the squared multiple correlation and squared cross-validity coefficient to violations of multivariate normality. *Applied Psychological Measurement, 6*, 185–200.

Drasgow, F., & Hulin, C. L. (1988). Cross-cultural measurement. *Interamerican Journal of Psychology, 21*, 1–24.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143–165.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. (Doctoral dissertation, Illinois Institute of Technology, 1993). *Dissertation Abstracts International, 54-04*, 2266B.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hanisch, K.A. (1992). The Job Descriptive Index revisited: Questions about the question mark. *Journal of Applied Psychology, 77,* 377–382.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hofstede, G. (1980). *Culture's consequences.* Beverly Hills: Sage.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296–309.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translation. *Journal of Applied Psychology, 67,* 818–825.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood: Dow Jones–Irwin.

Junker, B., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 31–61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32,* 261-276.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

Landar, H. J., Ervin, S. M., & Horowitz, A. E. (1960). Navaho color categories. *Language, 36,* 368–382.

Leung, K., & Drasgow, F. (1986). Relation between self-esteem and delinquent behavior in three ethnic groups: An application of item response theory. *Journal of Cross-Cultural Psychology, 17,* 151–167.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, Netherlands: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Marin, G., Gamba, R. J., & Marin, G. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology, 23,* 498–509.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-squared test of item bias with estimated and known person parameters. *Applied Psychological Measurement, 11,* 161–173.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG 3.* Mooresville: Scientific Software.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Richards, I. (1953). Toward a theory of translation: Studies in Chinese thought. *American Anthropological Association, 55* (Memoir 75). Chicago: University of Chicago Press.

Roznowski, M. (1989). An examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology, 74*, 805–814.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Segall, D. O. (1983). *Assessment and comparison of techniques for transforming parameters to a common metric in item response theory.* Doctoral dissertation, University of Illinois at Urbana-Champaign.

Smith, P. C., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally.

Smith, P. C., Sademan, B., & McCrary, L. (1992, May). *Development and validation of the Stress in General (SIG) scale.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Psychometrika, 27*, 229–239.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Stone, C. A. (2003). Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. *Educational and Psychological Measurement, 63*(4), 566–583.

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement, 60*(6), 974–991.

Triandis, H. C. (1972). *The analysis of subjective culture.* New York: Wiley.

Triandis, H. C. (1990). Cross-cultural studies of individualism and collectivism. In J. Berman (Ed.), *Nebraska Symposium on Motivation, 1989* (pp. 41–133). Lincoln: University of Nebraska Press.

Triandis, H. C. (1994). *Culture and social behavior.* New York: McGraw-Hill.

Triandis, H. C. (1995). *Individualism and collectivism.* Boulder: Westview Press.

12

إنشاء وتكييف وتقييم صدق اختبارات القبول بلغات متعددة: الحالة الإسرائيلية

مايكل بيلر
هيئة الاختبار التعليمية
ناعومي غافني وبنيناهاناني
المعهد الوطني للاختبار والتقييم

يؤثر اختلاف الأهداف التي تترجم الاختبارات من أجلها على عملية الترجمة وعلى دور كل من اللغات المستخدمة. يلزم نقاش مستقل حول الاعتبارات المتعلقة بملاءمة النسخة المترجمة لكل استخدام على حدة. إن أحد الاستخدامات الشائعة للاختبارات المترجمة هو تطبيق مقياس معياري موثوق مثل اختبارات معامل الذكاء IQ أو استبيانات الشخصية. وتستخدم هذه الاختبارات (مثل مجموعة تقويم كاوفمان للأطفال) [K-A-B-C] بشكل رئيس لتطوير المعدلات الإحصائية المحلية اللازمة لأغراض البحث العلمي. تتضمن عملية الترجمة في هذه الحالة الحد الأدنى اللازم من تكييف الاختبار بورتينغا، (1995)

إن التقويمات الدولية للتعليم (مثل دراسة التوجهات الدولية في الرياضيات والعلوم [TIMMs] هي أمثلة على الأبحاث عبر الدول التي لا تعتبرها أية لغة أو محتوى مسيطراً بل تحدد كل الدول المشاركة محتويات التقويم معاً مؤمنة بذلك قواسم مشتركة عظمى.

تترجم النسخة المتفق عليها إلى كل اللغات المشاركة (انظر غريسي، 2003 وهو مثال ممتاز) تكون المقارنة الرئيسة من هذه الحالة المقارنة بين الدول مع أن بعض الحالات (كما في كندا) المقارنة في بلد واحد بين مجموعات مختلفة.

هدف آخر لترجمة الاختبارات هو وضع سياسة انتقاء عادلة وصحيحة للمرشحين في مختلف المجموعات اللغوية للمتقدمين للقبول في معاهد التعليم العالي في بلد ولغة تعليم معينين. هذا الأمر شائع في البلدان التي تستقبل أعداداً كبيرة من المهاجرين (مثل: الولايات المتحدة وكندا وإسرائيل) فقد تؤثر عملية الترجمة في هذه الحالة وبشكل كبير على صدق قرارات فردية ذات نتائج خطيرة. إن استخدام درجات اختبارات القبول المجراة لكل المجموعات بلغة المصدر تؤدي إلى الخلط بين المنشأ المقاس ومستوى إتقان لغة المصدر؛ لذلك هناك حاجة لإيجاد الوسائل للتقليل من الخلط بين هذين المتغيرين وذلك على سبيل المثال بترجمة اختبارات القبول إلى لغات المتقدمين المختلفة وقياس إتقانهم للغة المحلية بشكل منفصل.

يجب معالجة أهداف الترجمة مع أخذ المجموعة السكانية المستهدفة بعين الاعتبار، حيث يختلف وضع هذه المجموعات من حالة لأخرى: في بعض الحالات يملك البلد الواحد أكثر من لغة رسمية واحدة (مثال: سويسرا، كندا، إسرائيل) وبالتالي فإن الاختبارات تترجم بشكل روتيني. وحتى في مثل هذه الحالات هناك اختلافات بين البلدان، فبعضها يؤمن للمجموعات اللغوية المختلفة نظاماً تعليمياً كاملاً بلغتها الخاصة (مثل سويسرا) فيما تؤمن بلدان أخرى نظاماً تعليمياً نصف منفصل (كندا، إسرائيل). ففي إسرائيل، على سبيل المثال، تشكل اللغة العربية لغة رسمية ثانية ويوجد نظام تعليمي منفصل للسكان الناطقين يمتد من الحضانة حتى نهاية التعليم الثانوي (باستثناء بعض الموضوعات التي تدرس باللغة العبرية). أما اللغة المعتمدة في التعليم العالي، فهي العبرية والعبرية فقط. هناك نظام تعليمي واحد لكلتا المجموعتين السكانيتين.

يشكل المهاجرون في البلدان المختلفة مجموعات سكانية إضافية مستهدفة في ترجمة الاختبارات. ويجب أن لا يعامل المهاجرون من بلد معين وكأنهم مجموعة متجانسة. فهم يتفاوتون في معرفتهم بلغتهم الأم وباللغة المحلية الجديدة وفي معرفتهم بالثقافة المحلية والنظام التعليمي (يتعلق ذلك بعمرهم عند الهجرة وطول المدة التي أمضوها في البلد الجديد، ومستوى انغماسهم في ثقافة البلد الجديد). لا تضمن ترجمة اختبار إلى لغة معينة (للأسباب المبنية أعلاه) مقارنة صحيحة حتى بين أفراد مجموعة لغوية واحدة تؤثر في الدرجات المختلفة في إتقان كل مجموعة سكانية للغة المصدر واللغة الهدف في طريقة ترجمة الاختبارات. فقد يكون من المفضل في بعض الحالات ترجمة تعبيرات معينة من الاختبار عوضاً عن ترجمته كاملاً. فعلى سبيل المثال، قد يفضل المهاجرون القدامى إلى إسرائيل (الذين درسوا في إسرائيل لعدة سنوات) أن يقدموا اختبارات القبول الجامعية باللغة العبرية مع ترجمة تعبيرات معينة إلى الروسية عوضاً عن أخذ الاختبار كاملاً باللغة الروسية.

يركز هذا الفصل على بعض المشكلات الكبرى المتعلقة بترجمة الاختبارات من جهة استخدام الاختبار وتحليل الدرجات. ويعالج هذا الفصل بشكل خاص المدى الذي يجب أن يصل إليه ترجمة أو تكييف أو تغيير النسخة الأصلية من الاختبار. كما يعالج تحديد معايير تقديم نوعية الترجمة ومقاربة معايرة الدرجات في النسخ اللغوية المختلفة.

تناقش أساليب معالجة هذه الموضوعات ويتم شرحها باستخدام النسخ اللغوية المختلفة من اختبار الدخول القياسي النفسي PET اختبار الدخول القياسي النفسي PET، هو اختبار أهلية مدرسية ويوضع ويجرى من قبل المعهد الوطني للاختبار والتقويم NITE ويستخدم بالإضافة إلى شهادة قبول لاتخاذ قرارات القبول في كل الجامعات الإسرائيلية ومعاهد التعليم العالي الأخرى. أما شهادة القبول فهي توضع بناء على تقييم المدرسة واختبارات أداء خارجية تجرى على نطاق وطني. يقيس

PET قدرات معرفية ومدرسية مختلفة في محاولة لتقدير النجاح المستقبلي في الدراسات الأكاديمية. ويتألف PET من ثلاثة اختبارات فرعية مختلفة متعددة الخيارات، المحاكمة اللغوية (V)، المحاكمة الكمية (Q)، والإنجليزية كلغة أجنبية (E) لا يوجد تصحيح للتخمين في وضع درجات الاختبار، بل يشجع الممتحنون على التخمين عندما لا يعرفون الجواب الصحيح. (انظر بيلر، 1994 للحصول على وصف تفصيلي لـ (PET لقد واجه الذين يقومون بالقياس النفسي والمخططون عند وضع سياسة القبول في جامعات إسرائيل مشكلة إيجاد الأساليب المثلى للتنبؤ بالنجاح الأكاديمي للمتقدمين غير الناطقين بالعبرية (بالإضافة إلى الناطقين بالعبرية) في معاهد التعليم العالي (حيث لغة التعليم هي العبرية). كان الهدف بعبارة أخرى تصنيف المتحنين في المجموعات اللغوية المختلفة على مقياس واحد يقيس معياراً مشتركاً في السياق الثقافي نفسه. تقرر بناء على ذلك إجراء PET باللغة التي يتقنها المتقدم بشكل أفضل لأنه كان يُعتقد أن هذا الأمر يؤمن للممتحن فرصة الأداء الأمثل. ويجري ترجمة PET حالياً إلى اللغات التي يتكلمها أغلب المتحنين: العربية، الروسية، الفرنسية، الإسبانية والإنكليزية. إن عملية الترجمة جهد مستمر، حيث تترجم سنوياً أربعة نماذج إلى العربية. نموذجان إلى كل من الروسية والإنكليزية ونموذج واحد إلى كل من الفرنسية والإسبانية (من أصل عشر نماذج إلى ثمانية عشر نموذجاً بالعبرية).

ينشر المعهد الوطني للاختبار والتقويم كتيّب معلومات يحتوي اختبارات مجراة سابقاً مع شرحها، وذلك لتعريف المتحنين بالـ PET والتأكد من فهم كل الأشخاص لمتطلبات كل مهمة في الاختبار بشكل فاعل. يترجم هذا الكتيّب إلى اللغات الخمسة المذكورة سابقاً. وهذا الإجراء مهم لأن المجموعات اللغوية المختلفة تتفاوت في مدى خبرتها السابقة مع الاختبارات متعددة الخيارات. من بين الـ 66731 شخصاً الذين تقدموا للاختبار (PET) عام 1998 على سبيل المثال، اختار 27 ٪ آخذ الاختبار من واحد من هذه اللغات (العربية15 ٪، الروسية 10 ٪، 2٪ من لغات أجنبية أخرى)

تطلب بعض المعاهد من الممتحنين الذين أخذوا PET بلغة أجنبية أن يأخذوا اختباراً إضافياً لإتقان اللغة العبرية(HP) وتوضع درجاته بشكل منفصل. إن النسخ غير العبرية من PET هي أساساً ترجمات للنسخ العبرية المجراة للناطقين بالعبرية، ولذلك فهي ذات بناء متماثل. يكون اختبار اللغة الإنكليزية الفرعي واحداً لكل النسخ اللغوية. ويترجم اختبار المحاكمة الكمية في العبرية، وأساس ذلك أنه يمكن مقارنة مواد الرياضة المترجمة بشكل مباشر بالمصدر. أما اختبار المحاكمة اللغوية فهو مترجم جزئياً فقط حيث تنتقى معظم البنود من مجموع البنود العبرية ويتم إنشاء بنود أخرى بشكل خاص للنسخ اللغوية المختلفة (مثل بنود المفردات والتعبير). نناقش في الفقرات التالية مسائل إجرائية وجوهرية في ترجمة PET، ويعطى اهتمام خاص لترجمة اختبار المحاكمة اللغوية.

عملية الترجمة

انتقاء نموذج اختبار للترجمة: بعض الاعتبارات

تترجم الاختبارات في اللغات الخمس (العربية، الروسية، الفرنسية، الإسبانية، الإنجليزية) من نماذج اختبار عبرية مجراة سابقاً. وهذا يعني أن المواد المنتقاة للترجمة ذات نوعية عالية بمعايير القياس النفسي. يجب مراعاة الاعتبارات التالية عند انتقاء النسخ العبرية للترجمة:

1- نوعية المعايرة: لتلافي حدوث مشكلات معايرة، نحاول تحديد نوعية المعايرة لنموذج مجرى سابقاً لممتحنين ناطقين باللغة العبرية مشابهين نسبياً من حيث توزيع القدرات للممتحنين باللغات الأخرى (المجموعة الهدف).

2- الجدارة: لقد وُجد جدارة الدرجة الكاملة في النسخ المترجمة كان قريباً جداً من جدارة النسخة العبرية الأصلية (انظر الجدول 12-1 أما في الماضي، فكانت جدارة اختبار المحاكمة اللغوية للمجموعة الناطقة بالعربية أقل من جدارة الاختبارات الفرعية المترجمة الأخرى (انظر الجدول 12-1 للقيم الحالية). كان السبب الرئيس هو أن اختبار المحاكمة اللغوية كان صعباً جداً للناطقين

بالعربية. ولزيادة جدارة اختبار المحاكمة اللغوية في هذه المجموعة تم إنشاء اختبار أسهل بانتقاء النصف الأسهل من المواد من نموذج عبري (ليستعمل لاحقاً في المعايرة) وإكمالها ببنود سهلة من مصرف الأسئلة.

3- الحفاظ على تواتر التعبيرات التقنية في نصوص القراءة والفهم: تم تحاشي نصوص القراءة والفهم التي تزخر بالتعبيرات التقنية (تعبيرات علمية، لغة قانونية، مصطلحات نفسية) عند انتقاء نصوص للترجمة. وذلك لأن هذه التعبيرات قد تكون مفهومة في لغة معينة دون الأخرى. بالإضافة إلى ذلك، فإن تواتر استخدام هذه التعبيرات يختلف باختلاف السياق الثقافي واللغوي، فلا يترجم نص ما يزخر بالكلمات الأجنبية إلى العربية لأن الممتحنين الناطقين بالعربية لا يصادفون مثل هذه الكلمات في المدرسة الابتدائية أو الثانوية.

4- السياق الثقافي: يجب أن يكون السياق الثقافي مألوفاً لكل الممتحنين. يجب عدم ترجمة نص قراءة فهم يحوي دلالات ثقافية محلية.

5- مراجعات الحساسية: تخضع مواد الاختبارات لمراجعات حساسية لتفادي اختبار مواد قد تكون مثيرة أو مؤذية في النسخ المترجمة. فعلى سبيل المثال لا نستعمل أي مادة تحتوي كلمة انتفاضة بالعربية بسبب الحساسية السياسية للكلمة. ويتم تفادي النصوص التي تناقش السياسة أو الدين أو الجنس، إلى ما هنالك.

جدول رقم 12-1

جدارة الدرجة الكاملة في النسخ المترجمة لاختبارات PET الثانوية

Language	*V*	*Q*	*E*	*PET*
Eebrew (65)	0.90	0.90	0.94	0.96
Arabic (23)	0.84	0.86	0.82	0.93
Russian (18)	0.86	0.88	0.92	0.94
French (9)	0.82	0.87	0.90	0.93
Spanish (9)	0.82	0.87	0.93	0.94
English (15)	0.90	0.90	0.96	0.96

ملاحظة: تظهر أعداد نسخ الاختبار بين أقواس.

مراحل عملية الترجمة:

تحقق عملية الترجمة شروط تكييف الاختبارات التي وضعتها الهيئة الدولية للاختبارات (هامبلتون، 1994، انظر أيضاً الفصل الأول من هذا الكتاب)، هناك أربع مراحل في عملية الترجمة:

1- الترجمة الأولية: يقوم مترجم مؤهل وخبير يتقن اللغتين ويعرف الثقافتين، خاصة الهدف، بترجمة النسخة العبرية الأصلية إلى اللغة الهدف. يناقش المترجم المشكلات التي تظهر في أثناء عملية الترجمة مع عالم القياس النفسي الذي يقود عملية الترجمة. تجرى حالياً ترجمتان مستقلتان إلى اللغة الروسية، وذلك بعد توصية من هامبلتون (في اتصال شخصي معه عام 1997). وتشير التجربة حتى الآن مع هذا الإجراء المكلف أنه يحسن نوعية المراجعة التي تجري في المرحلة التالية. ويجري الآن تقويم الجدوى الاقتصادية لهذا الإجراء الإضافي.

2- المراجعات المستقلة: توضع النسخ المترجمة لمراجعة نقدية من قبل عدة مراجعين ثنائي اللغة، يملك بعضهم خلفية قوية في الرياضيات والمنطق، فيما يبرع الآخرون في المحاكمة اللغوية. يقرأ مراجعون أمريكيون وبريطانيون النسخة الإنكليزية، فيما يقرأ مراجعون من عدة بلدان ناطقة بالإسبانية النسخة الإسبانية. يُطلب من المراجعين نقد النسخة المترجمة أولاً دون النظر إلى النسخة العبرية. وبعدها فقط يقارنون النسخة المترجمة بالنسخة العبرية الأصلية. ثم يُطلب منهم الانتباه بشكل خاص إلى دقة الترجمة، ووضوح الجمل، ومستوى صعوبة الكلمات، وسلاسة النص. يُقوّم كل مراجع كل بنود الاختبار ويتحرى أي تغيرات حدثت في المنطق الداخلي للبند، ويتأكد من أن كل بند ما زال يملك حلاً واحداً فقط لا غير وأن الإجابات المضللة مناسبة من حيث جاذبيتها. يناقش المترجم وعالم القياس النفسي تعليقات واقتراحات ويتخذان التعديلات المناسبة.

3- الترجمة الراجعة: يقوم خبير ثنائي اللغة، لم يسبق له رؤية النسخة العبرية، بترجمة النسخة المترجمة وبشكل شفهي إلى اللغة العبرية. وتُجرى هذه المرحلة شفهياً، وذلك يسمح بالنقاش بين عالم القياس النفسي والمترجم الراجع. تقارن الترجمة الراجعة مباشرة مع النسخة العبرية وتعدل المواد المترجمة عند الضرورة.

4- المراجعة النهائية قبل الإجراء الأولي للاختبار: تعطى النسخة المعدلة من الترجمة إلى ناطق أصلي باللغة الهدف. ويطلب منه حل الأسئلة دون النظر إلى الأصل العبري والتأكد من وجود حل واحد صحيح فقط لكل سؤال. يُقوّم عالم القياس النفسي الإجابات ويبحث عن الإجابات الخاطئة التي قد تنتج عن أخطاء في الترجمة.

المشكلات الخاصة بترجمة اختبار المحاكمة اللغوية:

إن فقرات المحاكمة اللغوية هي الأكثر إشكالية في الترجمة: لأن الكلمات والمفاهيم في إحدى اللغات قد لا تحافظ على المعاني والدلالات نفسها والاعتياد أو مستوى الصعوبة عند ترجمتها إلى لغة أخرى. تشكل العبارات والتعابير مشكلة شائعة حيث لا يمكن غالباً ترجمتها على الإطلاق. كما تختلف الكلمات في غناها اللفظي في مفرداتها. فعلى سبيل المثل، تزخر العبرية بالكلمات المتعلقة بالزراعة. ففي الإنكليزية يقال يقطف العنب أو يقطف الزيتون، وهكذا. أما في العبرية فهناك فعل مختلف لقطف العنب وقطف الزيتون وما شابه ذلك. كذلك يوجد في العبرية كلمات مختلفة لغسل الأرض أو غسل الصحون أو غسل الثياب. في حين يستعمل الناطقون بالإنكليزية فعل "غسل" لكل هذه الأعمال. تملك كل من العبرية والإنكليزية كلمة واحدة لـ "جمل"، في حين يوجد في العربية عدد كبير للتعبير عن الأنواع المختلفة من الجمال حسب مواصفاتها.

تشكل الترجمة إلى العربية مشكلة كبيرة. ففي حين أن العربية الفصحى هي نفسها لكل العرب في كل البلاد العربية، تختلف العربية المحكية بشكل كبير عن

العربية المكتوبة. كما تختلف من بلد إلى آخر وحتى من منطقة إلى أخرى من البلد نفسه. فكلمة "coat" تعني "معطف" بالعربية الفصحى، و"كبوت" باللغة العامية. أما "hat" فتعني "قبعة" في العربية الفصحى و"طاقية" بالعربية العامية. تتوخى الترجمة العربية للاختبار تجنب استخدام الكلمات العامية. وتستخدم عوضاً عن ذلك الكلمات الفصحى بالرغم من أنها أكثر صعوبة. فقد يواجه الممتحنون الذين يقرؤون الأدب هذه الكلمات، في حين قد لا يعرفها آخرون.

تناقش الفقرات التالية بعض مشكلات الترجمة التي تتراوح بين بعض أنواع البنود التي لا يمكن ترجمتها على الإطلاق، إلى تلك التي يمكن ترجمتها مباشرة أو بعد تعديل طفيف.

بنود استبدال الحروف. تستند هذه البنود إلى خاصية الصرف من اللغات السامية والتي لا تشاركها فيها اللغات الهندو -أوربية. وهي أن معظم المفردات في اللغة العبرية، أي أن كل الأفعال ومعظم الأسماء والصفات، تتألف من جذر + نهاية صرفية. تتألف بنود استبدال الحروف من أربع جمل تعدّل في كل منها كلمة واحدة بتغيير حروف جذرها إلى قالب ثابت هو عادة الحروف (PTI) يحل هذا القالب الثابت، في ثلاث من الجمل، محل الجذر نفسه، في حين يحل في الجملة الرابعة محل جذر مختلف وعلى الممتحنين تعيين هذه الجملة. يمكن استخدام هذا النوع من البنود في اللغة العربية لأنها لغة سامية. لكن البنود لا يمكن ترجمتها من العبرية، بل تكتب في العربية ويتم اختبار البنود الجديدة من هذا النوع قبل استخدامها في وضع الدرجات. لا يمكن استخدام هذا النوع من البنود في النسخ اللغوية الأخرى.

الكلمات والتعابير. لا يمكن ترجمة الكلمات والتعبيرات من العبرية. بل تكتب مباشرة باللغة الهدف. وقد وُجد أنه من الضروري والمجدي اقتصادياً في نسخة الاختبار العربية اختبار هذه البنود قبل استخدامها لأغراض التصحيح.

التشبيهات. هذا النوع من البنود هو الأصعب في الترجمة. حيث يتعلق بمعاني ودلالات كلمات إفرادية والعلاقة بين أزواج من الكلمات. يندر إيجاد كلمات تملك المعنى الدقيق نفسه والدلالات نفسها مستوى الصعوبة نفسه في لغة أخرى. يجب عند ترجمة التشبيهات الحفاظ بما يمكن من الدقة على العلاقة بين كلمتي كل زوج من الكلمات مع الإبقاء على مستوى صعوبة المفردات. يصمم التشبيه الأصلي عادة لاختبار إتقان مفردات اللغة العبرية، بالإضافة إلى القدرة التحليلية. وغالباً ما تكون البنية المترجمة أكثر سهولة.

إتمام الجمل. تستلزم هذه البنود هذه العلاقات المنطقية واللفظية في جملة مركبة. وتصعب ترجمة بنود إتمام الجمل. فمن أجل تأليف جملة طبيعية وانسيابية في اللغة الهدف، يلزم غالباً تغيير بنية الجملة مما يؤثر على موقع الكلمات الناقصة في الجملة المترجمة. ويجب على المترجم أن يتأكد من أن الإجابات المضللة تنتج جملاً صحيحة نحوياً وإعرابياً بحيث يعتمد اختيار الجواب الصحيح على المنطق الداخلي وليس على تلميحات بنيوية ونحوية. كما يجب الحفاظ على مستوى اللغة (عادية، رسمية، أدبية) وتعقيدات الكلمات المفقودة وعدد الفراغات.

تظهر المشكلات مثلاً في العربية حيث لكل اسم جنس نحوي قد لا يطابق جنس مقابله العبري. كما أنه يوجد في العربية شكلان من الجمع، الجمع الأكثر من بندين (الجمع) والجمع لبندين (المثنى) حيث يصرف الفعل بناءً على ذلك. كما تبدأ الجمل في العربية بفعل حين تبدأ في العبرية باسم. تستدعي كل هذه المشكلات عدة تعديلات في بنية البند وتعقد مهمة ترجمة بنود إتمام الجمل.

عندما تتغير بنية الجملة في اللغة الهدف، قد تحتوي الجملة ثلاثة فراغات عوضاً عن أربعة في النسخة العبرية الأصلية. ولا يوجد سبب مسبق لعدم استخدام مثل هذه البنود، أو إذا وُجد أن هذا البند في تحليل المعايرة أسهل بكثير من مقابله في العبرية، فسوف يلغى من المعايرات اللاحقة.

المنطق. يجب ترجمة بنود المنطق بعناية ودقة. يجب على المترجم أن يحاول الحفاظ على كل العناصر المنطقية في البند العبري مع الحفاظ على البنية الموجودة نفسها في البند الأصلي. يجب الانتباه إلى السياق فيما إذا كان حقيقياً أو خيالياً وتعديل الأسماء ووحدات القياس (كيلومترات أو أميال) بحيث تكون المصطلحات المستخدمة مألوفة لدى الممتحنين.

في محاولة للحفاظ على البنية الأصلية لفقرات المنطق الأصلية والنسخة المترجمة (مثل الحفاظ على النفي أو النفي المضاعف أو أحرف الربط مثل فقط.... إلخ (يلزم أحياناً تغيير البنية بحيث يكون التركيب الإعرابي في اللغة الهدف صحيحاً. أحد الأمثلة على ذلك أن التركيب الإعرابي للجملة العبرية "all p`s are not q" غامض في الإنكليزية. وهكذا فإن مقولة بالعبرية مثل "All birds of prey are not green" والتي تعني "أنه لا يوجد طير جارح واحد أخضر اللون"، لا يمكن ترجمتها مباشرة إلى الإنكليزية بل يجب تغييرها كالتالي "No birds of Prey are green"وتظهر صعوبة مماثلة عند ترجمة الجملة نفسها إلى الفرنسية. ففي الفرنسية، لا يمكن أن يأتي نفي بعد "كل" وبالتالي تصبح الترجمة "Aucun oiseau predate n`est vert"

القراءة والفهم. يتم التأكيد في ترجمة النص على النقاط التالية: دقة الترجمة، الحفاظ على سلاسة وغنى وأسلوب اللغة باستعمال مفاهيم مألوفة في اللغة الهدف واستعمال المصطلحات التي تظهر في النص بشكل متناسق.

إحدى الانتقادات التي توجه إلى الاختبار العابر للثقافات، هو أن النص المترجم لا يوصل المعنى نفسه ولا يحافظ على مستوى الصعوبة نفسه في النص الأصلي. لهذا، أنشئ في المعهد الوطني للاختبار والتقويم فريق خاص بهدف إيجاد نصوص خاصة للناطقين بالعربية، والذين يشكلون أكبر مجموعة من المتقدمين للاختبار غير الناطقين بالعبرية. تُكتب النصوص بالعربية ويتم تكييفها لمتطلبات اختبارات المعهد

الوطني للاختبار والتقويم. أحد الأسئلة التي يجب الإجابة عنها في المستقبل، -هو فيما إذا كانت الدرجات الناتجة عن اختبار يستخدم نصوص فهم مكتوبة أصلاً بالعربية يمكن مقارنتها مع الدرجات الناتجة عن نسخ الاختبار التي تستخدم نصوصاً" مكتوبة أصلاً باللغة العبرية.

تصحيح النسخ اللغوية

يتم تصحيح كل اختبار فرعي بشكل منفصل باستعمال صيغة قاعدة درجات. ويتم معايرتها على مقياس كان له في المجموعة المعيارية الأصلية (المتقدمين للاختبار الناطقين بالعبرية) 1984 متوسط قيمته 100 وانحراف معياري قيمته 20 أما الدرجة الكلية PET فهي المجموع المثقّل لدرجات الاختبار الفرعية ، (2V, 2Q + E) وله متوسط قيمته 500 وانحراف معيار قيمته 100 (لوصف أكثر تفصيلاً انظر بيلر، 1994).

تطبق الضوابط نفسها في تصحيح الاختبارات الفرعية للمحاكمة الكمية والإنكليزية في كل النسخ اللغوية (وذلك بافتراض أن الترجمة لا تغير معنى الأسئلة الكمية). يتم القيام بعملية معايرة شبيهة بالتي وصفها آنغوف ومودو (1973)، وتُستخدم في تصحيح اختبار المحاكمة اللغوية. تم تأسيس قاعدة مشتركة بين النسخة العبرية وكل من النسخ اللغوية الأخرى وذلك بانتقاء بنود ذات مؤشرات قياسية نفسية متشابهة وترتيب صعوبة متشابهة (باستخدام التقنيات البيانية دلتا) لمجموعتين من الممتحنين. عند وضع القاعدة المشتركة تطبق أساليب التعديل الخطي.

يظهر الجدول 12-2 المتوسطات والانحرافات المعيارية للنسخ اللغوية المختلفة لـ PET واختباراته الفرعية للعام الدراسي 1997/ 1998 المجموعات الناطقة بالفرنسية والإسبانية والإنكليزية صغيرة جداً، بحيث لا يمكن تعميمها. أما المجموعات الناطقة بالعربية والروسية فهي كبيرة ومستقرة عبر السنين بشكل كاف لتكون ممثلة لأكبر مجموعتي أقليات تتقدم للدخول إلى معاهد التعليم العالي في إسرائيل.

وهكذا، فإن أغلب التحليل والنقاش التاليين مبني على النسخ العبرية والعربية والروسية. كما ذكر سابقاً، يملك السكان الناطقون بالعربية في إسرائيل نظاماً تعليمياً منفصلاً لغة التعليم فيه هي العربية. يظهر التفاوت بين المجموعات الناطقة بالعبرية وتلك الناطقة بالعربية في الصفوف الأولى، كما أظهرت عدة تقويمات وطنية للنظام التعليمي. كما يظهر مستوى الأداء الأعلى بعض الشيء في الرياضيات لدى الناطقين بالعربية نسبة إلى التحصيل اللغوي في سن مبكرة.

من اللافت للنظر أن الفرق الأكبر في الأداء بين المجموعتين الناطقتين بالعربية والروسية والمجموعة العبرية، هو في اختبار اللغة الإنكليزي الذي لا يخضع للترجمة.

الجدول رقم 12-2

المتوسطات والانحرافات المعيارية لصور لغوية متعددة لاختيار PET والاختبارات الفرعية المصاحبة له في العام 1997/1998.

		Total Score		*Verbal Reasoning*		*Quantitative Reasoning*		*English*	
Language	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Hebrew	48,897	554	101	108	20	111	19	111	23
Arabic	9,949	431	85	86	16	92	19	82	16
Russian	6,366	512	101	92	18	106	18	95	22
French	511	521	84	99	16	104	19	112	17
Spanish	363	480	82	90	14	96	17	108	22
English	645	552	106	100	21	107	21	131	23

جودة النسخ المترجمة

بالإضافة إلى عملية الترجمة الدقيقة التي وُصفت سابقاً، تستعمل المعايير الكمية التالية لتقويم جودة النسخ المترجمة: الأثر التفاوتي للتخمين، تحليل البنود، الأداء التفاوتي للبنود ذات الجدارة العالية، تكافؤ البنية، المصداقية والانحياز في

التنبؤ بالدرجات. تتأثر هذه المعايير بالترجمة وبعوامل أخرى متنوعة متعلقة بثقافة المجموعة وتوزيعها الثقافي. تُعطى اللغة الروسية اهتماماً خاصاً لأنها لغة أكبر مجموعة مهاجرين في إسرائيل (15٪ من سكان إسرائيل).

الأثر التفاوتي للتخمين: بحثت دراسات أجراها غافني وميلاميد (1994) الظاهرة التالية: بالرغم من توجيه الممتحنين إلى التخمين إذا لم يعرفوا الجواب الصحيح، فإن 75 ٪ - 93٪ منهم فقط (حسب الاختبار الفرعي) أجابوا على كل بند في PET لقد افتُرض أن المجموعات اللغوية المختلفة قد تظهر أنماطاً مختلفة من السلوك من ناحية التخمين. كان من المتوقع، على سبيل المثال، أن يكون الناطقون بالإنكليزية أكثر اعتياداً على اختبارات الخيارات المتعددة، أن يتبعوا تعليمات الاختبار بدقة. أما الناطقون باللغة الروسية من جهة أخرى، فقد يكونون أقل ميلاً للتخمين. وذلك لقلة اعتيادهم على هذا النوع من الاختبارات. كما افتُرض وجود تأثير لمعرفة الجمهور العام للاختبارات ذات الخيارات المتعددة.

أوحت النتائج أن الناس المختلفي الخلفيات الثقافية مختلفون في ميلهم إلى التخمين. وُجد تأثير للمجموعة وتأثير الاعتياد. في عام 1984 كان المتحنون الناطقون بالروسية والعربية والفرنسية يميلون إلى إغفال بنود أكثر من الناطقين بالعبرية والإنكليزية والإسبانية. وفي عام 1987 (أي بعد أربع سنوات من إجراء PET كان المتحنون بالروسية يميلون إلى إغفال بنود أكثر من كل المجموعات الأخرى. لكن نسبة البنود المغفلة انخفضت بشكل ملحوظ في كل المجموعات، حيث أصبح الاختبار أكثر اعتياداً، وتحضير الاختبار أكثر شيوعاً. أوصت دراسة في عام 1994 بالتركيز على تعليمات الاختبار، خاصة بين أفراد المجموعات الأكثر ميلاً لتجنب التخمين.

تحليل البنود والأداء التفاوتي للبنود:

تاختبار نوعية كل بند مترجم (من حيث مستوى الصعوبة ودرجة التحيز). كما ياختبار بالإضافة إلى ذلك، الأداء التفاوتي لكل بند مترجم وذلك بمقارنة المتحنين

الناطقين بالعبرية مع غير الناطقين بها (يشير تعبير الأداء التفاوتي للبند إلى الملاحظات البسيطة في أن البند يُظهر خصائص إحصائية مختلفة للمجموعات المختلفة، وذلك بعد استبعاد فروق القدرات بين المجموعات). إذا كانت الخصائص الإحصائية لبعض البنود المترجمة سيئة، تتم مراجعة هذه البنود والبحث عن أسباب فشلها. بعد ذلك يتم اتخاذ قرار حول إدخاله في القاعدة المشتركة المستخدمة للمعايرة.

قام كل من غافني ويهو شافات (1993) باختبار الأداء التفاوتي للبنود في اختبار المحاكمة اللغوية لثلاثة نماذج روسية من PET باستخدام تقنية الخط البياني دلتا المقترحة من قبل آنغوف (1927). ولوحظ أن أكبر أداء تفاوتي للبنود كان في التشبيهات. وأصغر أداء تفاوتي للبنود كان في بنود المنطق وإتمام الجمل. هذه النتائج مشابهة لتلك التي وجدها آنغوف وكوك (1988) للممتحنين الناطقين بالإنكليزية والإسبانية في اختبار التقويم المدرسي (SAT) وذلك باستثناء بند المنطق غير الموجود في اختبار التقويم المدرسي اللغوي. وقد أظهرت بنود القراءة والفهم أداءً تفاوتياً للبنود أكبر نسبياً منه في دراسة آنغوف وكوك.

وقد فسر كل من آلالوف وهامبلتون وسيرسي (1999) وسيرسي وآلالوف (2003) نتائج بحثهم عن علاقة الأداء التفاوتي للبنود مع نوع البند (باستعمال طريقة مانتل - هانسل) ووضعوا نظريات حول أسباب ومصادر الأداء التفاوتي للبنود. قام هؤلاء بتحليل ثلاثة نماذج من النسخ العبرية والروسية لـPET تعكس النتائج مدى المشكلات المتعلقة بترجمة الأنواع المختلفة للبنود اللغوية كما وضحنا في الفقرة السابقة. فقد وجدوا أن 42 من أصل 125 بنداً أظهرت أداءً تفاوتياً عبر اللغات. كانت التشبيهات الأكثر إشكالية، حيث أظهرت 65 ٪ منها أداء تفاوتاً. وقد أظهر الممتحنون الناطقون بالروسية نتائج أفضل في معظم هذه البنود من الناطقين بالعبرية. كما أن نسبة كبيرة من بنود إكمال الجمل (45٪) أظهرت أداءً تفاوتياً. ولكن في هذه الحالة كان أداء المجموعتين اللغويتين متقارباً. وتم الطلب من

المترجمين أن يفكروا في أسباب الأداء التفاوتي لكل بند، فكانت الأسباب الرئيسة المقترحة هي تغييرات صعوبة الكلمات، تغيرات في الصيغة، تغيرات في العلاقة الثقافية وتغييرات في المحتوى.

الجدارة:

تقدر الجدارة الداخلية لكل اختبار فرعي وللعلامة الكاملة بشكل روتيني لكل نسخة لغوية. يظهر الجدول 12-2 متوسط معاملات الاتساق الداخلي (KR-20) للاختبارات الفرعية الثلاثة وللعلامة الكاملة في النسخ اللغوية المختلفة من PET قد تفسر جدارة النسخ الأجنبية الأولى نسبياً، بمشكلات متعلقة بالترجمة. ولكن الجدارة الداخلية لا تُحدد فقط بجودة بنود الاختبار وجودة الترجمة، بل أيضاً بالتباين الحقيقي في مجموعة الممتحنين. تظهر الخبرة المكتسبة في المعهد الوطني للاختبار والتقويم أنه في كثير من المهارات تم الخلط بين نوعية الترجمة وفروقات الأداء الحقيقية. فعندما تتفاوت مجموعتان في القدرة، فإن ذلك بحد ذاته قد يخلق فروقاً في الجدارة وقابلية المقارنة و الأداء التفاوتي للبنود. فعندما تكون البنود صعبة لمجموعة معينة، تضعف جدارة الاختبار في هذه المجموعة. فعلى سبيل المثال: كان متوسط الجدارة لاختبار المحاكمة اللغوية في النماذج العربية الخمسة الأولى التي وضعت بين عامي (1984 و1989) 0.68 (بيلر وغافني 1995). ولرفع هذه الجدارة، تم إنشاء اختبار محاكمة لغوية بشكل خاص للنسخة العربية حيث استعمل فيه بنوداً أكثر سهولة فارتفع متوسط الجدارة الناتج عن 23 بنداً إلى 0.84 وعلى الرغم من أن جدارة هذا الاختبار الجديد ارتفعت، إلا أنها تسبب خطأً معدلاً أكبر من الاختبار السابق.

وعلى الرغم من أهمية هذه النتائج، إلا أن الجدارة هي شرط لازم وليس كاف لتطوير الاختبار. إن مصداقية النسخ المترجمة المختلفة هي التي تعطي المبرر لاستخدام الدرجات الناتجة عنها.

تكافؤ البنية:

للتأكد من أن النسخ المترجمة لـ PET تقيس البنية نفسها التي تقيسها النسخة العبرية، استخدم آلالوف وباستاري وهامبلتون وسيرسي (1997) تحليل العامل الاستكشافي والقياسي متعدد الأبعاد، وتحليل العامل المؤكد لتقويم التكافؤ البنيوي لاختبار المحاكمة اللغوية في نسختين عبرية وروسية. وقد حللوا بالتحديد أربعة من مجالات المحتوى الخمسة: التشبيهات، النطق، القراءة والفهم، وإتمام الجمل. تضمن التحليل 41 بنداً. وقد وُجد في التحليلات التي أجريت على النسختين أن بنية PET كانت متشابهة عبر النسختين اللغويتين بالنسبة لهذه البنود.

الصدق:

تُختبر عملية الانتقاء بشكل روتيني وذلك باختبار الصدق التنبؤي لـ PET ومقارنته بمعيار متوسط الدرجات النسبي (GPA) في نهاية السنة الدراسية الجامعية لأولى، وعند إتمام الدرجة العلمية الأولى. يركز هذا الفصل على المقارنة بين نتائج الممتحنين الناطقين بالعبرية والروسية. أما نتائج دراسة الصدق وانحياز الاختبار، بالنسبة للنسخة العربية منPET، فيمكن مراجعتها في بيلر، غافني وهاناني (1999).

الصدق التنبؤي للنسخة الروسية مقارنة بالنسخة العبرية:

تتأثر الصدق التنبؤي للنسخة الروسية بدرجة أقل بعوامل مثل الفوارق الكبيرة في المقدرات بين المجموعتين الناطقتين بالعبرية والعربية. لهذا تناقش الأبحاث المتعلقة بالترجمة الروسية بشكل أوسع من الأبحاث المتعلقة بالنسخة العربية.

قامت دراسة أجراها مؤخراً غافني وبرونر (1998) بحساب الصدق التنبؤي لعلامة PET ومقارنتها مع نظيرتها في المجموعة الناطقة بالعبرية. كانت المتنبئات في هذه الدراسة هي PET واختباراته الفرعية الثلاثة (V, Q, and E) ودرجة القبول (Adm) كانت درجة القبول في هذه الدراسة مرتكزة على ثقلين متساويين في PET ودرجة تحصيل معطاة إما في المدرسة الثانوية (باغروت) أو في برنامج تحضره للممتحنين الذين لم يدرسوا في مدرسة ثانوية إسرائيلية. وقد احتوت هذه

الدراسة متنبأً إضافياً هو علامة اختبار إتقان اللغة العبرية (HP) الذي يُجرى لكل المتحنين غير الناطقين بالعبرية. ثم تم حساب معاملات الصدق للمتنبئات المختلفة للطلاب الذين بدؤوا دراستهم الجامعية بين العامين 1992- 1996 لمعيار في متوسط الدرجات النسبي للسنة الأولى (FGPA) ومتوسط الدرجات النسبي للسنة الثالثة (TGPA) أجريت التحليلات لكل قسم جامعي ضمن كل جماعة شرط أن تضم خمسة طلاب على الأقل من كل من المجموعتين اللغويتين. تم إصدار النتائج لـ463 قسماً حقق الشروط السابقة لـ (FGPA و83 قسماً حقق الشروط لـ TGPA).

يظهر الجدول 12-3 عدد الطلاب والمتوسط والانحراف المعياري لمختلف المتنبئات والمعايير.

الجدول رقم 12-3

المعدلات والانحرافات المعيارية (داخل الأقواس) لدرجات كل من المتنبئ والقاعدة (العبرية والروسية) لكل من TEPA, FGPA

Language	*PET*	*V*	*Q*	*E*	*Adm*	*HP*	*FGPA*	*TGPA*
FGPA								
Hebrew	600	118	117	118	101	-	80	-
N = 55,434	(60)	(13)	(13)	(16)	(6.8)		(8.8)	
Russian	561	111	116	100	99	93	73	-
N = 7,313	(52)	(12)	(12)	(16)	(6.2)	(16)	(11.9)	
TGPA[a]								
Hebrew	590	116	116	116	101	-	82	84
N = 6,612	(57)	(12)	(13)	(15)	(6.2)		(6.7)	(6.9)
Russian	540	108	112	96	98	87	76	81
N = 1,011	(54)	(13)	(12)	(16)	(6.1)	(14.4)	(8.4)	(8.3)

كان الفارق الأكبر بين المجموعتين في اختبار الإنكليزية (E)،(P) حيث أظهر المتحنون الناطقون بالعبرية أداء أفضل. لم يوجد أي فارق في اختبار المحاكمة

الكمية (Q) أما الفارق في اختبار المحاكمة اللغوية (V)، فكان في الوسط. كما وُجد فارق طفيف لصالح الممتحنين الناطقين بالعبرية في علامة القبول (Adm) مما يوحي بنمط معكوس في الفوارق في درجة التحصيل (غير المعايرة) بالمقارنة مع PET كان الفارق في معيار (FGPA) مماثلاً للفارق في PET، ومن اللافت للانتباه أن الفوارق تقلصت في (TGPA) مقارنة ب (FGPA).

يظهر الجدول 12-4 معاملات الصدق للمجموعتين اللغويتين عبر معدلها في كل الأقسام، وتصحح الارتباطات الملاحظة (بين قوسين) لمحدودية المدى. كان معدل معاملات الصدق لكل من درجة القبولPET في مجموعة (FGPA) متماثلاً في كل من المجموعتين العبرية والروسية عبر كل مجالات الدراسة.

الجدول رقم 12-4

صدق التنبؤ (صححت الارتباطات لمحدودية المدى) لكل من درجة PET: درجة القبول، واختبار الكفاءة العبري (HP) والمعدل العام (GPA) في نهاية السنة الجامعية الأولى (FGPA) والسنة النهائية (TGPA) لكل من متحدثي الروسية والعبرية من الممتحنين (يظهر الترابط الخام داخل الأقواس).

Language	*PET*	*V*	*Q*	*E*	*Adm*	*HP*
FGPA						
Hebrew	.39	.32	.36	.24	.48	
	(.26)	(.21)	(.26)	(.12)	(.37)	
Russian	.35	.26	.30	.29	*	*
	(.27)	(.16)	(.21)	(.24)	(.38)	(.23)
TGPA						
Hebrew	.44	.36	.37	.29	.54	
	(.20)	(.16)	(.17)	(.12)	(.28)	
Russian	.45	.35	.38	.35	*	*
	(.26)	(.17)	(.19)	(.22)	(.33)	(.20)

معاملات الصدق لكل من درجة القبول PET في مجموعة (FGPA) متماثلاً في كل من المجموعتين العبرية والروسية عبر كل مجالات الدراسة.

لكن نمط معاملات الصدق في الاختبارات الفرعية لـ PET كان مختلفاً في المجموعتين اللغويتين حيث إن Q كان الأعلى مصداقية في المجموعة الناطقة بالعبرية. في حين كان E الأكثر صدقا في المجموعة الناطقة بالروسية مما قد يشير إلى أن V لا يقيس تماماً البنية نفسها في اللغتين. وهذا يعود إما إلى ترجمة تكييف الاختبارات أو بسبب محتوى الاختبار بعينه الذي تم اختياره أصلاً للمجموعة الناطقة بالعبرية. يبدو أن عوامل عدة تحدد مصداقية اختبار في مجموعة معينة، وجودة الترجمة واحد منها فقط. يمكن أن تعزى المصداقية العالية نسبياً لاختبار الإنكليزية (E)، (P) في المجموعة الناطقة بالروسية إلى متغيرات معدلة لم يتم تحريها في هذه الدراسة. فمثلاً يمكن أن يكون الطلاب الذين هاجروا إلى إسرائيل من مدن كبرى ذات نظام تعليمي متطور وحصلوا على فرص أفضل لتعلم اللغة الإنكليزية من أولئك الآتين من بعض المدن البعيدة التي تفتقر إلى نظام تعليمي حديث ومتطور. كما أنه من الممكن أن يكون بعض المتحنين الناطقين بالروسية قد هاجروا إلى إسرائيل منذ سنين عديدة ودرسوا في النظام التعليمي الإسرائيلي لعدة سنوات قبل تقديم PET وينعكس هذا في درجة اللغة الإنكليزية ودرجتهم المعيارية.

انحياز الاختبار:

قام المعهد الوطني للاختبار والتقويم بإجراء بحث لتقصي وجود انحياز في الاختبار للممتحنين الناطقين بالروسية (غافني، برونر، 1998) ويشير التعبير "انحياز" إلى أخطاء منهجية في الصدق والتنبئي أو صدق المفهوم المتعلق بمجموعة معينة من المتحنين. تم استخدام الطرق المستمدة من التعريفات الواردة في دارلينغتون (1971) والنقاش الذي قام به لين (1984) لتحري الانحيازات في المتنبئات المختلفة. ويجب تفسير النتائج المتعلقة بمتنبأ واحد بحذر، وذلك لتأثير إخراج متنبأ ما من معادلة راجعة فيها فروق موجودة أصلاً بين المجموعات لينو ورتس (1971).

الانحياز في اختبار الممتحنين الناطقين بالروسية:

تتألف العينة الأولى من 55.434 ممتحناً ناطقاً بالعبرية، و7.313 ممتحناً ناطقاً بالروسية. بدؤوا دراستهم في إحدى السنوات 1992- 1996 وكانت درجاتهم في FGPA وPET متوفرة. تم إشراك ست جامعات إسرائيلية وما مجموعه 463 قسماً في الدراسة. كانت درجة القبول (Adm) متوفرة لعينة قريبة تتألف من 26.875 ناطقاً بالعبرية و3.478 ناطقاً بالروسية. أُجري التحليل لهذا المتنبأ في 259 قسماً فقط (غافني، برونر 1998). يظهر الجدول 12-5 عدد حالات الانحياز المهمة التي تم الكشف عنها في كل الأقسام واستخدام درجة القبول وAdm، PET كمتنبئين.

جدول رقم 12-5

عدد حالات الانحياز المهمة التي تم الكشف عنها في كل الأقسام واستخدام درجة القبول وPET،Adm كمتنبئين

Predictor	*Bias Against Russian Speakers*	*Bias Favoring Russian Speakers*
FGPA		
PET	3.0	2.0
V	0.5	1.5
Q	1.0	7.0
E	1.5	0.0
Adm	1.0	9.0
TGPA		
PET	8.0	5.5
V	2.7	5.5
Q	2.7	5.5
E	11.0	0.0
Adm	2.7	0.0

لم يظهر انحياز واضح عند استخدام PET كمتنبئ منفرد: ففي 3٪ من الـ463 حالة كان هناك مؤشر واضح على الانحياز ضد الممتحنين الناطقين بالروسية، وكان الانحياز في 2 ٪ لصالحهم. وقد وُجدت نتائج مماثلة لـ "V"و "Q" (الاختبارات

المترجمة) مع ميلها إلى مبالغة النتائج المتوقعة لـ FGPA للمجموعة الناطقة بالروسية. وقد وجد ميل معاكس في اختبار الإنكليزية E أما في درجات القبول، فكان هناك في حوالي 10 ٪ من الـ 259 قسماً مؤشرات واضحة على الانحياز، وغالباً لصالح الممتحنين الناطقين بالروسية.

لا تعطى النتائج المذكورة آنفاً في إجابات واضحة على انحياز التنبؤ الممكن حدوثه نتيجة الترجمة. فقد أظهر اختبار الإنكليزية E، وهو غير مترجم، انحيازاً ضد الناطقين بالروسية، بينما أظهر اختبار المحاكمة الكمية Q، وهو غير متأثر بالترجمة نسبياً، انحيازاً لصالح هذه المجموعة. أما اختبار المحاكمة اللغوية الأكثر تأثراً بالترجمة فقد أظهر نتائج لا تشير إلى أي مشكلات أكيدة في الترجمة. يمكن الاستنتاج أنه إذا نجحت عملية الترجمة في الحفاظ على مستوى الصعوبة نفسه في النسختين اللغويتين، وإذا كان معنى ما يتم قياسه متشابهاً قدر الإمكان، فلا يجب توقع أي انحياز نتيجة للترجمة بحد ذاتها.

إن أحد الانتقادات الرئيسة لدراسات الانحياز التقليدية هي أنها تبالغ في تنبؤات الدرجات المعيارية لمجموعات الأقليات، إن النقلة إلى كلية يشكل فيها الغالبية السكانية القسم الأكبر من الطلاب يفرض متطلبات أكثر على طلاب الأقليات منه على طلاب الغالبية السكانية. ولهذا يتوقع المرء أن تختفي المبالغة التنبئية في السنة الجامعية الثالثة. تم اختبار عينة فرعية مؤلف من (6.612) ممتحناً ناطقاً بالعبرية و (1.011) ممتحناً ناطقاً بالروسية تتوفر درجاتهم فيPET وFGPA لاختبار هذا الافتراض. وقد ضم هذا الاختبار 83 قسماً. كانت درجة القبول (Adm) متوفرة لعينة فرعية مؤلفة من (2.687) ممتحناً ناطقاً بالعبرية و(338) ممتحناً ناطقاً بالروسية. توفر التحليل لهذا المتنبأ في 37 قسماً فقط. يظهر الجدول 12-5 عدد الأقسام التي تم فيها اكتشاف انحياز واضح ضد ولصالح الممتحنين الناطقين بالروسية. وبشكل عام، عند استخدام TGPA كمعيار، ظهر نمط مماثل لـ FGPA في انخفاض بسيط من ميل المتنبئات إلى انحياز لصالح المجموعة الناطقة بالروسية كما هو متوقع.

الاستنتاجات

يجب مقاربة قضية ترجمة الاختبارات مع التفكير بالأبعاد المختلفة هدف الترجمة، المجموعة السكانية المستهدفة ونوع ومحتوى الاختبار. يمكن ترجمة اختبار ما لأغراض الأبحاث حيث تكون الأهمية الأولى للفروق بين المجموعات. كما يمكن استخدام الترجمة لاتخاذ قرارات فردية مهمة مثل القبول في الجامعات. إن متطلبات جودة الترجمة أعلى في الاستخدام اللاحق من السابق.

تم وصف عملية ترجمة PET وهو اختبار يستعمل في القبول في التعليم العالي في إسرائيل من العبرية إلى العربية والروسية والفرنسية والإسبانية والإنكليزية بشكل مفصل مع توضيح المشكلات المتأصلة في الترجمة (بشكل خاص في الفقرات اللغوية). تمت مراجعة جودة الترجمة بتطبيق طرائق نوعية وكيفية متنوعة، وذلك لإيضاح الجوانب المختلفة في عملية الترجمة للنسخ اللغوية المختلفة. تم اتخاذ خطوات متعددة لتأمين جودة الترجمة:

(أ) استثمار جهد كبير في المراجعة النوعية للترجمة مع استخدام مترجمين مستقلين في بعض الحالات.

(ب) اختبار أنماط الإجابات والميل التفاوتي للممتحنين إلى التخمين في بنود الخيارات المتعددة.

(ج) اختبار تحليل البنود والأداء التفاوتي للبنود.

(د) اختبار الجدارة وعلاقتها بمستوى قدرات المجموعة.

(هـ) تحري الصدق التنبئي لمعيارين (معدل الدرجات النسبي للسنتين الدراسيتين الجامعية الأولى والثالثة).

(و) تحليل انحياز الاختبار التنبئي للمجموعات الفرعية المختلفة.

إن المجموعة الواسعة من التحليلات المقدمة في هذا الفصل والمتعلقة بـPET، والنسخة المترجمة تعطي الكثير من المعلومات حول تعقيد مسألة تساوي وتكافؤ

اختبارات القبول المترجمة. يؤخذ هذا الرأي القائل إنه يجب تبني وجهة نظر أوسع لعدالة الاختبار وتكافؤ النسخ. عند تقويم جودة الاختبارات المترجمة في سياق قرارات فردية مهمة، يجب القيام ببعض الصدق التنبئي وانحياز الاختبار، بالإضافة إلى التحليلات الأخرى الأكثر شيوعاً مثل اختبار الأداء التفاوتي للبنود. إن مسألتي صدق الاختبار وانحيازه ضروريتان لوضع الأساس لاستخدام نتائج الاختبار، ليس أقل أهمية من الاختبار نفسه في سياق اختبارات القبول.

تظهر النتائج المقدمة في هذا الفصل أنه عند تطبيق عملية ترجمة جيدة لـPET، يمكن إنتاج مجموعة من الاختبارات المترجمة المتكافئة البينية والجدارة الصحيحة والعادلة نسبياً. ولكن حتى عندما تُتخذ كل هذه الخطوات للتأكد من جودة الترجمة وتساوي الدرجات، يظل الحصول على نسختين واحدة أصلية وأخرى مترجمة أمراً مستحيلاً. ويظل الخيار الآخر الذي يقضي باختبار الناطقين بغير العبرية حلاً أقل عدالة. كما يجب اعتبار عوامل أخرى مثل التأثيرات الاقتصادية والفوائد المتوقعة (أي تقدير الكلفة المرافقة لتحسين العملية الحالية).

شكر

نرغب بشكر مايا بار- هيدل، غيرشون بن- شاكار، روث فورتس، شافا كاسلوجيري ليفينسون لملاحظاتهم العميقة، وشمول برونر لمساعدته في تحليل المعطيات.

*****

المراجع

Allalouf, A., Bastari, B., Hambleton, R. K., & Sireci, S. G. (1997). *Comparing the dimensionality of a test administered in two languages* (Laboratory of Psychometric and Evaluative Research Rep. No. 319). Amherst, University of Massachusetts, School of Education.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

Angoff, W. H. (1972, August). *A technique for the investigation of cultural differences.* Paper presented at the meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Rep. No. 88-2). New York: College Entrance Examination Board.

Angoff , W. H., & Modu, C. C. (1973). *Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Rep. No. 3). New York: College Entrance Examination Board.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13*(2), 12–20.

Beller, M., & Gafni, N. (1995). Translated scholastic aptitude tests. In G. Ben-Shakhar & A. Lieblich (Eds.), *Studies in psychology* (pp. 202–219). Jerusalem: The Hebrew University, The Magnes Press.

Beller, M., Gafni, N., & Hanani, P. (1999, June). *Constructing, adapting, and validating admissions tests in multiple languages.* An invited paper presented at the International Conference on Adapting Tests for Use in Multiple Languages and Cultures, Georgetown University, Washington, DC.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8,* 71–82.

Gafni, N., & Bronner, S. (1998, April). *An examination of criterion-related bias for Hebrew- and Russian-speaking examinees in Israel.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Gafni, N., & Canaan-Yehoshafat, Z. (1993). *An examination of deferential item functioning for Hebrew and Russian-speaking examinees in Israel.* Paper presented at the 24th Annual Conference of the Israeli Psychological Association, Ramat-Gan.

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation, 20,* 309–319.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 224–229.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33–47.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*(1), 1–4.

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment, 11*(3), 140–146.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–166.

13

التكييف عبر الثقافات للاختبارات التربوية والنفسية

بيتر ف. ميريندا
جامعة رود آيلند

إن مجال الاختبارات والتقييم التربوي والنفسي مليء بالممارسات المغلوطة العديدة التي نجم عنها كثير من العواقب الخطيرة. لسوء الحظ، فإن علماء النفس ومن يتفق معهم من الخبراء قد فشلوا في إدراك خطورة هذه المشكلات بشكل عام والتي سببها سوء استخدام الاختبارات. من الملاحظ أنه بين الأخطاء الرئيسة (التي تأتي في المقام الأول) المرتكبة، هي تطبيق وتفسير أدوات التقييم بشكل غير منظم على الممتحنين الذين يُسبب لهم معوقات اللغة وعوامل ثقافية أخرى إبطال صدق الاختبار. إن أوضح مثال على تلك الأخطاء هو سوء استخدام الأدوات عبر الثقافات في التفسير الخاطئ للدرجات التي تم الحصول عليها من تطبيق التقاويم غير المناسبة للثقافة المتلقية.

خلال الخمسين سنة الماضية التي ظهرت خلالها هذه الاستخدامات الخاطئة والتي بدأت على نطاق واسع في أنحاء العالم كافة، قام العديد من علماء القياس النفسي وعلماء النفس الدوليين بالإضافة إلى المؤلفين، بالكتابة بشكل مكثف عن هذه الأخطاء كي يتم تجنبها وعن الطرق المناسبة لتعديل أدوات القياس ثقافياً (انظر: بيهلينغ ولو، 2000، هامبلتون و دي يونغ، 2003، فان دي فيفر و لونغ،

1997). ونذكر بعضهم أيضاً، بيري (1997)، برسلين، لونر، ثورنديك (1973)، كرونباخ و درنث (1972)، كيسنجر (1994) أولمدو (1979)، بورتينغا (1995)، وسبيلبيرغ، دفيلس، وبوكلك (1994).

إن المشكلات الأخرى، التي غالباً لا تتم مناقشتها في الموضوعات المتعلقة بنقل أدوات التقويم عبر الثقافات، هي من غير ريب الافتراضات البسيطة في الثقافات المتلقية. وهي (أ) إن الأدوات ذات الجدارة والصادقة والمناسبة في ثقافة ما، يمكن تكييفها بسهولة لثقافات أخرى. (ب) التغاضي عن حقيقة أن هذه الأدوات نفسها صحيحة حسب القياس السيكولوجي في ثقافة المنشأ كما هو مفترض (ميريندا، (1994وهذا الأخير ينطبق على بعض الأدوات المكيفة الجديرة والمستخدمة بشكل واسع عبر الثقافات ميريندا (1995).

إن هذا الفصل يقدم بياناً مختصراً (عن السنوات الأولى من القرن العشرين) عن تكييف الاختبارات بهدف التغلب على عوائق اللغة في الولايات المتحدة والخارج. يقدم الجزء الأكبر من هذا الفصل التجارب والمشكلات التي واجهت المؤلف طوال 40 عاماً، كان يقوم خلالها بإنشاء أدوات تقويم تربوية ونفسية في الولايات المتحدة والقيام بأبحاث شاملة لتكييفهم إلى ثقافات أخرى. يوجد أيضاً موجز للإجراءات السليمة في تكييف الأدوات (انظر هامبلتون و باتسولا، 1999، فان دي فيفر و لونغ، 1997).

التكييف في الاختبارات التربوية والنفسية عبر الثقافات

إن إحدى الممارسات الخطيرة وعديمة التأثير في حقل الاختبار الثقافي والتقويم النفسي في نصف القرن الماضي، والمستمر حتى الآن، هو النقل السيئ لقياس الأدوات من ثقافة إلى أخرى أو إلى ثقافة ثانوية (ميريندا 1993). إن الممارسة السيئة تتضمن استعارة اختبار من ثقافة واستخدامه في ثقافة أخرى. لاحظ أن التأكيد هو على الاستخدام لا على تكييفه إلى ثقافة أخرى. يكون هذا

بالقيام بالترجمة الحرفية أكثر منها بالترجمة المتقدمة أو الترجمة الراجعة. وكما هو واضح في هذا الكتاب، هناك حاجة إلى الإثبات التجريبي أيضاً. بالرغم من أن ترجمة الاختبار يمكن أن يتبعه محاولة إعادة القياس، هناك ممارسة شائعة أخرى وهي ببساطة القيام بتفسير الدرجات المبنية على أساس القياسات الأصلية. لا يعطي تكييف البنود إلى الثقافة المتلقية إعادة قياس الإجراءات الإدارية والدرجات، وإثبات قياس البنية، غالباً الاعتبار المطلوب في تكييف الاختبار للاستخدام في لغة مختلفة حسب مبادئ قياس سيكولوجية معتمدة (الجمعية التربوية الأمريكية للأبحاث، الجمعية الأمريكية النفسية، المجلس القومي للقياسات التربوية 1999).

موجز عن التكييف الثقافي للاختبارات

من الممكن تتبع المحاولات الأولى لإقامة أدوات تقويم للتغلب على نقص المهارة في لغة الثقافة/ الثقافة الثانوية والتأثيرات المتحيزة على العوامل الأخرى التي تؤثر على الأداء إلى أوائل القرن العشرين. قامت إحدى المحاولات الرئيسة في الولايات المتحدة خلال الحرب العالمية الأولى مع جيش بيتا (Beta). أجرى تطوير الاختبار علماء نفس جرى تعيينهم في فيلق مهندسي التصريف الصحي، وحدة جيش خاصة تحت إمرة الميجور ر.م.يركيس، الذين قاموا بإدارة عملية تطوير الاختبار.

صُمم اختبار الجيش بيتا (Beta) لتقييم المجندين المتطوعين الذين كانوا إما أميين أو يعرفون لغة أجنبية. يتكون الاختبار من صور وأشكال تتطلب قلة أو عدم معرفة باللغة الإنكليزية. ويتضمن الاختبار أيضاً اختبارات أداء ثانوية مثل المتاهات، تصميم المجموعات وإنشاءات هندسية.استخدم اختبار بيتا بشكل كامل في الجيش بين عامي 1917 و 1918 جنباً الى جنب مع اختبار الأبجدية اللغوي للجيش.

تاريخياً، فإن اختبارات ألفا وبيتا هي اختبارات تُطبق على مجموعات ويُعمل بها في الولايات المتحدة بشكل واسع. في وقت مبكر، منذ عام 1904 طُورت لوحة استمارة لاختبارات أداء لا تتطلب معرفة باللغة الإنكليزية واستخدمت على مستوى أصغر.

لأسباب خاصة ذات دلالات تاريخية، ربما بسبب شهرة الشخص الذي أنشأ واستخدم تلك اللوحة، طُورت لوحة الاستمارة من قبل ر. س. وودورث. في عام 1904 طبق وودورث ذلك الاختبار في المعرض العالمي في سانت لويس على أطفال لا يتكلمون اللغة الإنكليزية. لاحقاً في دراسة عن الاختلافات العنصرية، استخدمت وطبعت النتائج مع لوحة الاستمارة واختبارات أداء أخرى (وودورث 1910).

قبل دخول الولايات المتحدة الحرب العالمية الأولى (1917- 1918)، أدرك هـ.أ. نوكس (1914) الخطورة والعجز في اختبار المهاجرين (الذين يعانون من علل عقلية) في جزيرة أليس باستخدام أدوات طورت باللغة الإنكليزية في الولايات المتحدة. بدأ بتطوير اختبارات لا تتطلب إجابات باللغة الإنكليزية. إن اختبار لوحة الاستمارة التي طورها نوكس كانت لوحة استمارة تقرير الإجابة "Casuist Form Board". كانت تلك اللوحة معقدة أكثر من لوحة استمارة سكوين. وتتكون بشكل رئيس من دوائر خشبية وحواجز دائرية. طور نوكس أيضاً لوحة استمارة حيث يجب على الممتحنين استخراج نسخة عن نموذج مطبوع على مجموعة من المكعبات. في ذات العقد (1910-1919)، جرت عدة تطورات مماثلة. امتحن هيلي وفيرنالد (1911) الأطفال في معهد الأحداث للاضطرابات العقلية في شيكاغو الناطقين بلغة أجنبية واحدة، أو الذين لديهم إعاقة تمنعهم من الإجابة في الاختبارات العادية الموجودة في ذلك الوقت.

أحد الاختبارات غير الشفهية التي استخدموها، كانت اختبار إكمال الصور (Ebbinghaus picture Completion).

في عام 1911 بدأ رودولف ودونالد باترسون في تطوير مقياس اختبار الأداء المعروف والمستخدم بشكل واسع. تم نشر هذا المقياس (الذي كان الاختبار الأول للاستخدام الميداني) وأصبح متوفراً للاستخدام العملي بعد عدة سنوات (بنتز وباترسون1917) يحتوي ذلك المقياس على 12 اختباراً منها 8 اختبارات شفهية. من

ضمن اختبارات الأداء في المجموعات، كان ترتيب الدوائر البلاستيكية، حل الأحجيات، اختبار إكمال الصور واختبار محاكاة.

في العقد الثاني من القرن العشرين، كان ستانلي بورتوز يطور اختباره الشهير، اختبار المتاهة، الذي تم نشره للمرة الأولى في منتصف العقد (بورتوز، 1915). في بداية العقد التالي، طور كوس (1920) اختبار تصميم المجموعات (تصميم تجريبي يُقسم فيه أفراد التجربة إلى فئات متعددة، كل فئة تمثل مجموعة متجانسة بالنسبة إلى هدف التجربة) الذي بقي حتى الآن كجزء من مجموعة قياسات أداء واختبارات ذكاء طورت في الولايات المتحدة، وجرى استعارتها لثقافات أخرى. اتبعت غودنوف الاختبارات غير الشفهية (1926) وطورٌ اختبار " ارسم رجلاً" لقياس ذكاء الأطفال بين عمر 3-13 سنة. طُلب من الأطفال أن يرسموا صورة رجل، "أفضل صورة يمكن رسمها". كانت الدرجات حسب عدد الأجزاء المهمة للرجل مثل العينين، الأصابع، الأنف، الفم. إلى ما هنالك، الموجودة في رسم الطفل. لم يكن للنوعية الفنية أي أهمية. قادت المعايير التي اعتمدت على 400 طفل إلى استنتاج العمر العقلي(MA) ومعامل الذكاء (IQ) فيما بعد. طُورت تلك المعايير كمقياس لتشخيص علم النفس المرضي للأطفال الشاذين.

التطورات خارج الولايات المتحدة

كانت المحاولة الأولى المهمة خارج الولايات المتحدة لتطوير أداة غير متحيزة ثقافياً، حيث الكتابة أو اللغة الشفهية له جزء صغير جداً أو حتى منعدم، هي اختبار «ريفن الجدول التقدمي» (Raven Progressive Matrices) (ريفن1938) اليوم، وكما استخدم في بريطانيا منذ 50 عاماً، يستمر الاختبار بكونه الأداة للتقويم الشفهي الأكثر شعبية الذي طُبق في العالم، حيث يوجد الكثير من برامج تقويم المقاييس.

في القارة الأوربية، تم تطوير اختبار شفهي فريد في إيطاليا خلال الحرب العالمية الثانية. تم تصميم اختبار "بيديني" لتوزيع الانتباه لقياس مقدرة الشخص

على تركيز انتباهه على مهمة يجب عليه القيام بها بسرعة كبيرة، ثم يحول انتباهه إلى مهمة معكوسة (ميجليورينو1947). إن اختبار بيديني مفيد في تشخيص الضرر الدماغي أو الضرر في الجهاز العصبي. يعد ترميز ألوان التنقيط والخطوط العامل الأساسي في اختبار الأداء، والعائق الأول لأداء ناجح هو على الأغلب بدني (عمى الألوان، رؤية ضعيفة). جرت دراسة اختبار بيديني في الوقت الحاضر في الولايات المتحدة بغرض تطوير مقياس وصدق الأداة لاستخدامها في الثقافة في أمريكا الشمالية (ميريندا ودليوناردو 1992).

استخدم اختبار شفهي آخر في إيطاليا (اختبار"G" كالفي 1970) لتقويم القدرة الاستدلالية للمفاهيم لشخص ما. قدّم هذا الاختبار كونه مقياس عامل وحيد للذكاء العام الذي يمكن أن يفسَر كمعامل ذكاء تام، أو مبادئ تحليل "سبيرمان" للذكاء (g) الذي ما زال متوفراً للاستخدام. تم تصميم اختبار كالفي للاستخدام في إيطاليا، حيث لعبة الدومينو معروفة ومجموعات الألعاب شائعة في كل البيوت من البلاد. كانت الفكرة الضمنية هي إمكانية استخدامه عالمياً في كل ثقافة. بالرغم من وجود مجموعة من الإرشادات الكتابية مع الاختبار، إلا أنه يمكن أعطاؤها بالإشارة.

يتألف الاختبار من سلاسل متتالية للوجوه على قطع الدومينو. كل واحدة تعطي دلالة على تابع منطقي. على الممتحن أن يختار من عدة خيارات لقطع مختلفة الوجه الصحيح الذي يكمل تسلسل الفكرة. إن الكتيب المرافق والمنشورات التقنية صادقة للترجمة إلى لغة أي ثقافة متلقية. ولكن يبقى من الضروري للثقافة المتلقية أن تعيد إثبات خاصية القياس السيكولوجي.

إن البنود في اختبار كالفي لولبية، أي أنها تقدّم إلى الممتحنين حسب درجة الصعوبة المتنامية. يقدم بندان سهلان جداً في أول سلسلة البنود، ثم تصبح البنود أكثر صعوبة وتعقيداً وهكذا حتى آخر الاختبار.

تطور الاختبارات في الولايات المتحدة خلال الحرب العالمية الثانية

خلال الحرب العالمية الثانية، طور علماء النفس في الجيش الأمريكي اختبارين منفصلين للجنود المتطوعين. (أ) اختبار الأبجدية للجيش، وهو اختبار تجميع كتابي لتصنيف المتطوعين والمطلوبين للخدمة العسكرية المتعلمين والناطقين باللغة الإنكليزية. (ب) اختبار بيتا للأميين وللرجال الناطقين بلغة أجنبية. خلال الحرب العالمية، لم يكن هناك أي محاولة على مستوى كبير لخدمة الجنود والبحارة الذين كانوا يعانون من إعاقة بسبب عدم وجود المهارة اللغوية. كان للقوات البحرية وللقوات الجوية برامجها الخاصة للخدمات. طورت القوات البحرية اختبار مجموعة أشياء مترابطة ومصنفة للمستخدمين المجندين. تتكون المجموعة من اختبار مصنف عام (GCT)، اختبار حساب (ARITH)، اختبار كتابي (CLER)، و اختبار ميكانيكي (MECH) بينما طورت القوات الجوية مجموعة من اختبارات الأهلية واختبارات المهارات النفسية الحركية لاختيار متمرنين لثلاث وظائف لطاقم الطائرة، ربان الطائرة و المدفعي والملاح. طور الجيش اختبار التصنيف العام للجيش (AGCT) الذي يعطي قياس درجات عام يرتكز على مقياس ذي متوسط 100= (M) انحياز مقياس 20 = (SD).

يقوم هذا الاختبار بقياس القراءة، المفردات، وتقدير الحساب، الاستنتاج الحسابي والعلاقات المكانية. بالإضافة إلى استخدامه لتصنيف المجندين إلى اختصاصين مهنيين، واختيار الذين يحرزون درجات قياس أكبر من متوسط الدرجات بدرجة على الأقل (+ 120) للتدريب ليصبحوا ضباطاً. لم تكن هناك أية محاولة أثناء قيام الحرب لتطوير نسخة اختبار الجيش بيتا الذي استخدم في الحرب العالمية الأولى. لم يعط الجيش أي شرح عن عدم تطوير اختبار مواز لاختبار بيتا في الحرب العالمية الثانية. إن المعطيات التي قامت بجمعها وتحليلها هيئة أبحاث العلوم الاجتماعية وتم نشرها تحت رعايتها، تظهر بوضوح الحاجة إلى مثل

هذا الاختبار. إن اختبار التصنيف العام للجيش صنف المستخدمين إلى خمس فئات حسب الدرجات النهائية. كانت أغلب المجموعات الحاصلة على الدرجة الدنيا من السود أو الأشخاص ذوي الأصول اللاتينية (ستوفر وزملائه، 1950).

على أية حال، بعد انتهاء الحرب، وقّع قسم أبحاث المستخدمين في مكتب الجيش في واشنطن DC، على عقد اتفاقية مع مجلس البحث التربوي من كامبردج، مستشوست، لتطوير بدائل عديدة للاختبار الشفهي ليقوم مقام اختبار بيتا القديم. كان رئيس مجلس البحث التربوي، فيليب رولون، وهو أستاذ في القياسات والإحصائيات في جامعة هارفارد، كلية التربية. وقد عمل كثير من طلاب الدكتور رولون معه لإتمام شروط العقد. في عام 1952 تم تسليم الجيش اختبار الذكاء اللفظي (STI). تم تصميم تلك الأداة الشفهية لتقويم مفهوم قدرة الاستدلال للشخص. عند القيام بذلك الاختبار يتم تعليم الممتحن بالإشارة معنى إشارات خاصة عالمية، مجردة وملموسة. تم إنشاء ذلك الاختبار على مبدأ "التشبع الثقافي" بمواجهة " ثقافة محايدة". تم ربط مجموعات مركبة لأشكال هندسية مستوية مثلاً، مربعات، دوائر، معينات ومثلثات مع أشكال سوداء مظللة لحيوانات، أو أنثى إنسان في أوضاع متحركة أو ثابتة. على سبيل المثال تقفز أو في وضع الاستلقاء. لقد تم نشر الاختبار للاستخدام في الجيش في الولايات المتحدة، ولكنه لم ينشر للاستخدام المدني مثل اختبار التصنيف العام للجيش (AGCT). أراد رولون أن يسمح له الجيش باستخدام استمارة ينشرها مجلس الأبحاث التربوية بعد تقاعده من جامعة هارفارد (1967)، ولكن لسوء الحظ، توفي في ربيع 1968 وذلك قبل أن ينشر هو ورفاقه الطبعة المدنية للاختبار.

التطور بعد انتهاء الحرب العالمية الثانية

ابتدأ استخدام أدوات التقويم التربوي والنفسي (الاختبارات، المقاييس والبيانات) في حقل المعرفة والانفعالات بعد انتهاء الحرب العالمية الثانية بوقت قصير. وقد تم تطوير تلك الأدوات بشكل أساسي في الولايات المتحدة، بريطانيا

وفرنسا حيث توسعت استخدام التقاويم التربوية والنفسية بشكل سريع. وذلك في الخمسينيات والستينيات حين تسارع نمو علم النفس كمجال للعلوم في الولايات المتحدة. وبدأت الممارسة المهنية وتوسعت محفزة الاهتمام الكبير في عملية التقويم. أما في بلدان العالم الثالث، فقد بدأ علماء النفس والاختصاصيون باستعارة أدوات التقويم المطورة والمقاسة في بلدان أخرى بسبب نقص الخبراء ارتفاع التكاليف. لسوء الحظ، كان الإجراء الأساسي لنقل الاختبار من ثقافة إلى أخرى هو الترجمة اللغوية. وتستمر هذه الممارسة حتى الوقت الحالي إلى حد كبير. في بعض الأحوال يعطى قليل من الانتباه أو لا يعطى أي انتباه إلى الاختلافات الثقافية التي يجب اعتبارها عند تعديل أو تغيير البنود في الاختبار. كما لا يعطى أي اهتمام بإعادة القياس، المعايرة، وإجراء تكافؤ الدرجات والمسودة الأصلية للاختبار.

لإيضاح الفكرة السابقة، يجب الإشارة إلى أن خطأ الممارسة يحدث في البلاد الأقل تطوراً أو حتى في البلاد المتطورة. في أوربا، خاصة في بلدان البحر المتوسط. أما في بلدان شمال الألب وفي البلاد الإسكندنافية، فيبدو أنهم مطلعون على الإجراءات، ويمارسون مبادئ القياس السيكولوجي بشكل صحيح. أما في هولندا فهم يقدمون عدداً كبيراً من علماء القياس السيكولوجي المدربين جيداً والذين يمارسون على نطاق واسع في العالم. على سبيل المثال، عند مراجعة ثلاثة إصدارات لمجلة في القياس النفسي (Psychometrika) كانون الثاني 1998 حتى أيلول 1999) تبين أن 13 من أصل 25 مقالة منشورة قد كتبها علماء قياس سيكولوجي من هولندا.

ازداد في الولايات المتحدة التنوع في الثقافات والأعراق في الخمسينيات والستينيات. ولكي يستطيعوا تطبيق تشريع الإجراء الإيجابي الأول 1946، كان على ناشري الاختبارات الحصول على تقويم تربوي ونفسي لأدوات تم نشرها سابقاً في اللغة الإنكليزية وترجمت إلى لغات الأقليات. كان هذا صحيحاً بالنسبة للمجموعات من

الأصول اللاتينية، خاصة الأطفال في كوبا وبورتوريكو. على كل حال، هناك الكثير من المشكلات القياسية في تلك الطبقات باللغة الإسبانية ولغات أجنبية أخرى، لأن أكثرها تُرجم حرفياً ولم يجر تكييفه ثقافياً حسب الإجراءات القياسية الصحيحة.

لتأكيد هذا الخطأ وممارسته المستمرة في الولايات المتحدة من قبل ناشري الاختبارات، على المرء أن يدرس الفهارس الحالية التي تقوم بالدعاية لبيع مطبوعات لاختبارات اللغة الإنكليزية في لغات أجنبية. ما عدا بعض الاختبارات، فإن كل المطبوعات باللغة الأجنبية، هي ترجمة حرفية منقولة عن الأداة الأصلية.

إجراءات لتكييف الأدوات

إن الإجراءات المناسبة لتكييف الأدوات عبر الثقافات قد تم إيجازها في هذا الفصل، وجرى وصفها وشرحها في فصول أخرى من هذا الكتاب وفي مطبوعات أخرى (فان دي فيفر ولونغ 1997). هذه الإجراءات الشاملة والطويلة كانت السبب للنفقات المادية والوقت والجهد الذي أُنجز لتكييف سليم. ليس من الطبيعي أن تتطلب الإجراءات، لكي تكون ذات فاعلية، سنوات عديدة من العمل الجاد. هذه الجهود المشتركة التي قام بها اختصاصيون من البلدين المنشأ والمتلقي. الإجراءات هي كما يلي:

1- إن الخطوة الأولى هي الأخذ بالاعتبار الأدوات، التقنيات أو الوسائل المناسبة للتكييف. يجب أن يتضمن هذا الرأي تقويماً موضوعياً في صدق مواصفات القياس للمعايير في ثقافة المنشأ. يجب تجنب الخطأ، وهو افتراض صدق الأدوات لأنها ببساطة، استخدمت بشكل موسع.

2- قبل البدء بخطوة الترجمة يجب مراجعة البنود واستمارات الإجابة لاختبار طرق المجموعات الثقافية المحددة أو المجموعات الدولية.

3- في خطوة الترجمة، الترجمة الباكرة، والترجمة الراجعة، يجب الانتباه للتأكد من أن المترجمين خبراء في اللغتين ويعملون بشكل منفرد في المرحلتين. كما يجب الانتباه إلى صدق ترجمة المتغيرات في اللغة واللهجات لتأثيرها على الترجمة من حيث المفهوم والمعنى. وهذا شيء صحيح خاصة للغات اللاتينية. مثلاً اللغة الإسبانية في بورتوريكو مقابل الإسبانية في المكسيك، اللغة البرتغالية في أمريكا الجنوبية مقابل البرتغالية في البرتغال، أو اللغة الفرنسية مقابل الفرنسية في كندا.

4- يجب دراسة كل بند حسب إمكانية تكييفه إلى الثقافة المتلقية. من الثابت وجود بعض البنود التي لا يمكن نقلها مباشرة. يمكن تعديل هذه البنود وإصلاحها أو إهمالها ووضع بدائل، قبل البدء بتطوير نسخة تجريبية للاستخدام في الدراسة الاستطلاعية الأولى.

5- القيام بدراسة استطلاعية، حيث يجري استخدام النسخة التجريبية حسب الأعراف الثقافية، الممارسات، العادات، إلى ما هنالك للحصول على نماذج ذات تطابق كان في المجموعات التي تطبق عليها المقاييس.

6- عند تحليل المعطيات التي تم الحصول عليها عند تطبيق النموذج التجريبي، وهي خطوة مبكرة، يجب دراسة بنية وعينة (العامل أو العناصر) الأداة ومقارنتها مع بنية وعينة الأداة الأصلية. انظر غيرل، (2000) في هذه المرحلة، يمكن أن يقرر الباحثون ضرورة إعادة بعض الخطوات السابقة قبل الاستمرار.

7- إذا كان القرار أنه بالإمكان القيام بخطوة نحو تطوير النسخة المكيفة، فإن الخطوة التالية هي القيام بالتحليل الإحصائي الضروري والمطلوب لإثبات خصائص القياس السيكولوجي في الأداة المكيفة ثقافياً. في هذه المرحلة، وكحد أدنى، يجب حساب درجة الثبات الداخلية للبنود وصدق الدرجات والمعايير.

8- أما الخطوة الأخيرة في تطوير النموذج التجريبي، فهي القيام بدراسات للبنية ولصدق المعايير تنسجم مع الأغراض الأخرى التي تنوي الأداة استخدامها في الثقافة المتلقية؛ لأن نتائج بحث صدق الأداة الإيجابية سوف تؤمن مستخدمين محتملين وناشرين/ موزعين جداً في الثقافة المتلقية مع الثقة الحقيقية، وهي أن أداة مطورة جديدة مكيفة ثقافياً جاهزة للاستخدام الميداني.

إن هذه المجموعة من الخطوات الثمانية تتماشى مع المخطط الدولي لهيئة الاختبارات وتكييف الاختبارات الذي قُدم في الفصل الأول من هذا الكتاب (ووسع من قبل هامبلتون وباتسولا 1999).

نماذج من إجراءات استخدمت في تطوير الأدوات المكيفة اختبرها المؤلف خلال ٢٨ عاماً (- ١٩٦٧- ١٩٩٥)

جرى استخدام الأدوات المكيفة ثقافياً بشكل أولي في بحث أو أكثر من مشاريع الأبحاث عبر الثقافات وهي: (أ) تعيين المواهب في الدول النامية، 1977-1967، (ب) مطابقة الأولاد الصغار مع مشكلات التعليم، (1995-1975) و(د) الإدراك الحسي العام للقادة العالميين. 1964-1990، جرى ترجمة الأدوات لجذر من اللغة الإنكليزية إلى لغة البلد الذي يجري فيه البحث.

فيما يلي وصف مختصر لهذه الأدوات:

تحليل القوة الموجهة للنشاط (Activity Vector Analysis):(AVA) هي قائمة صفات أنشأها ولتر ف.كلارك. أصبحت AVA جاهزة للاستخدام. في 1948 استخدمت بشكل مبدئي في مجال الصناعة والتجارة في الولايات المتحدة (كلارك 1956). يحتوي الاختبار بشكله الأولي على 81 صفة للسلوك والتي تعطي درجات حسب أربع مقاييس وثلاث "أربعة عوامل" لصور صممت لقياس إدراك قدرات ذات الشخص الأساسية. الذات الاجتماعية والذات المركبة. تم إنشاؤها على

نظرية الذات الأساسية لـ لكي (لكي1945) وعلى النظرية الانفعالية لـ و.م. مارستون (1928 مارستون و.م.، كينغ ومارستون، ي. هـ 1931).

قياس المهارات (Mos)(Measurement of skills) : مجموعة من ثمانية اختبارات أنشأها ولتر ف.كلارك وزملاؤه. صمم هذا الاختبار لمساعدة العاملين في دائرة الموظفين في الاختيار السليم، والتصنيف وفي تعيين المهمات للموظفين. كانت الأدوات اختبارات قصيرة (الوقت المحدد 5 الى 7 دقائق) جرى توسيعها لمضاعفة الصدق لدقيقة من الاختبار. في هذه المجموعة من اختبارات الأداء المعرفي العملي، كانت المهارات المقاسة: (أ) المفردات، (ب) الأرقام، (ج) الأشكال (د) السرعة والدقة (هـ) التكيف (و) التفكير (ز) الذاكرة، و(ح) مهارة استعمال الأصابع (كلارك 1960).

مقياس التعريف بالطلاب في رود آيلند (Rhode Island Pupil Identifi cation Scale) (RIPIS):

صمم هذا المقياس للتعرف على الطلاب في الصفوف العادية (K-2) الذين يعانون صعوبات في التقدم المدرسي لعدة أسباب. يتألف المقياس من جزأين. يرتكز القسم الأول المكون من 22 بنداً على سلوك الطالب في الصف، والذي يمكن للمعلم مراقبته وإدراكه. ويعتمد القسم الثاني المكون من 19 بنداً على السجل المدرسي للأداء، والذي تسجله المعلمة في ملف كل تلميذ. هناك خمسة عوامل يتم الحصول عليها من القسم الأول للاختبار الأصلي: الإدراك الجسدي، تعاون حركي حسي، الانتباه، إدراك الذات وذاكرة الأحداث. تكون قيمتهم 67.5٪ من معامل التفاوت الكامل. وهناك أربعة عوامل من القسم الثاني للاختبار: الذاكرة لإعادة إخراج الرموز، التوجيه الثابت، الترتيبات المكانية المتتابعة، وذاكرة للرموز لعمليات الفهم. كانت قيمة هذا الاختبار 68 ٪ من معامل التفاوت الكامل (نوفاك، بونافنتورا وميريندا 1972/1979).

قوائم مسح المواهب في مشروع فلاناغان (Flanagan`s Project Talent Inventories):

استخدمت ثلاثة من قوائم فلاناغان في الأبحاث عبر الثقافات. وكانت كالتالي (أ) قائمة الاهتمامات، (ب) قائمة نشاطات الطالب، و(ج) قائمة فراغات تملأ بالمعلومات. كانت قائمة الاهتمامات تحتوي على 205 بنود تعالج 122 مهنة ونشاط يختار الطالب منها ما يهتم به (5 نقاط من مدرج مقياس الشخصية). وتحتوي قائمة نشاطات الطالب 150 بنداً حيث يجيب الطالب على (5 نقاط من مدرج مقياس الشخصية) للجملة التالية، "الأشياء التي أقوم بها وكيف أقوم بها، هل تصفني هذه الجملة". وتقدر إجابته من جيد الى جيد جداً. أما القائمة الفارغة للمعلومات فتتألف من 394 بنداً وفيها أسئلة منتقاة بدقة عن الخلفية الاجتماعية والخطط، طموحات طلاب في المرحلة المتقدمة في المدرسة. تنقسم هذه القائمة الى 7 أقسام، يستعلم القسم الأول (K=115) عن النشاطات التي شارك فيها الطلاب حتى ذلك الوقت. يتألف القسم الثاني (K=45) من أسئلة حول العائلة والمنزل. أما القسم الثالث والرابع فلهما علاقة بطبيعة الأعمال التي قام و بها والدا الطالب أو رب العائلة. ونشاطات أخرى قام بها أفراد العائلة. يتعلق القسم الخامس بالحالة الصحية للطالب (K=42) والجزء السادس بمخططات الطالب المستقبلية (K=75) وأخيراً يتعلق الجزء السابع (K=15) بالخطط المستقبلية للانتساب إلى الجامعة (فلاناغان وزملائه، 1964، ميريندا، ميجيليورينو1997).

تم اتخاذ الحيطة في كل مرحلة للتأكد من أن الترجمة كانت صحيحة للغة المجموعة التي تم تكييف الأداة لاستخدامها عليها. على سبيل المثال (أ) في تحليل القوة الموجهة للنشاط (AVA) جرت ترجمة الصفات لمختلف المجموعات الناطقة بالإسبانية. وحسب هذا تُرجمت الى الإسبانية القشتالية (الرسمية) للاستخدام في إسبانيا. اللهجة الإسبانية المستخدمة في كوبا/ بورتوريكو للاستخدام في الولايات المتحدة الشرقية، والإسبانية شيكانو المستخدمة في الولايات المتحدة الغربية

والمكسيك. استخدمت ذات العملية في تطوير اللغة الفرنسية. جرى التكييف الى الفرنسية الباريسية للاستخدام في فرنسا، الفرنسية الكندية للاستخدام في كندا، وفرنسية معدلة للاستخدام في السنغال. طُور شكلان من اللغة البرتغالية واحد منهما للاستخدام في البرتغال والثاني للاستخدام في البرازيل. جرى ترجمة مقياس التعريف بالطلاب الى لغات عديدة: الفارسية (إيران)، الماندرين الصينية (تايوان)، البولندية (بولندا)، الدنمركية (الدنمرك)، الفرنسية الكريبولية (هاييتي)، الإيطالية (إيطالية وسيسيليا). كما جرى ترجمة استبيانات فلاناغان الى الإيطالية الرسمية (توسكانا). بالرغم من أن هذه الأدوات قد تم استخدامها في مشروعات أبحاث كبيرة جرت في سيسيليا، فإن اللغة الإيطالية الرسمية هي اللغة الوحيدة التي تستخدم في المدارس في كل إيطالية. كان من الضروري اللجوء الى اللغة السيسيلية. بالرغم من كل الحذر الشديد أثناء القيام بمرحلة الترجمة - أثناء عملية التكييف، ظهرت كثير من المشكلات. وبعضها أكثر جدية من الأخرى. كان أكبرها وأكثرها كلفة المشكلة التي حدثت في تكييف مقياس التعريف بالطلاب (RI-PIS) في ايطاليا.

أجرى دراسة النسخة الإيطالية داميكو وزملاؤه (داميكو، ميريندا وسباراسينو، 1982 كان عامل الانتباه في الاختبار يتكون من ثلاثة أقسام: "الصعوبة في الجلوس ساكناً" #9، "الصعوبة في الوقوف ساكناً" #10، " ومدى انتباه قصير" #11.

كشف تحليل العناصر الرئيسة حسب تفاوت التناوب أن معظم العناصر تتداخل مع بنية المقياس في الولايات المتحدة. على كل حال، تم الكشف أيضاً عن نتيجة غريبة في عامل الانتباه. أخفقت البنود 9-10-11 في الربط مع المجموعة. ولكن بندين آخرين " البكاء" #15، "الإخفاق في تحمل التأنيب جيداً" #16 كانا ذوي ثقل في المجموعة. في اختبار RIPIS، في الولايات المتحدة، تظهر البنود 15-16 في عامل إدراك الذات على نحو متعاقب. أدت تلك المفارقات الجدية الى تكرار

الدراسة مع نموذج جديد مكون من 1.571 طفلاً في الصف k-2 من أربع مدارس في باليرمو/ سيسيليا. تم اكتشاف خطأ في ترجمة البنود 9-10-15 من اللغة الإنكليزية الى الإيطالية. كان المقصود من البنود 9-10 هو اتخاذه تلك الوضعية، ولكن الترجمة وجهت إجابة المدرسين الى مهمة الوقوف أو الجلوس بسكون. (وافق على تلك الترجمة أربعة مترجمين ثنائيي اللغة). كان الوضوح في المعنى المقصود ناجحاً في البندين 9-10 .

تم أخذ نموذج ثالث (N = 1.311) من باليرمو من أربع مدارس في الصف k-2 أصبح مقياس التعريف بالطلاب RIPIS باللغة الإيطالية الآن نسخة مكيفة ومعدلة من النسخة الأصلية (الأمريكية) (سيرسي، كارديكا و كانجيمي، 1992) وأصبح جاهزاً للاستخدام في 1995 وهو الآن يوزع من قبل أكبر دار نشر إيطالية Organizzazioni Speciali في فلورنسا.

في أثناء عملية تطوير نسخ من لغات أجنبية لاختبار تحليل القوة الموجهة للنشاط (AVA) كان من الضروري استبدال الصفات ببنية قواعدية أخرى. على سبيل المثال، حال، مصدر، وإلى ما هنالك. لم يكن ذلك لعدم وجود ترجمة حرفية للصفة، ولكن لأن الترجمة كان لها معنى مغاير في اللغة المقصودة. مثلاً كلمة (اجتماعي) في اللغة الإنكليزية تصبح في الفرنسية (يحب العيش ضمن مجموعة). وكلمة (يُعجب) تصبح في اللغة الإيطالية (يدرك الذات). بدأت الكثير من الأبحاث حول الصفات في AVA في 1957 في الولايات المتحدة، وفي عام 1967 بدأت في بلدان أجنبية حيث جرى تكييف قوائم الصفات بهدف القيام بالأبحاث.

بالإضافة الى التأكد من صحة الترجمة والمحتوى الثقافي في البند تبعه إجراء فعال في التكييف الثقافي لأداة التقويم، وهو تغيير إجراءات الاختبار الى الأفضل إذا كانت قابلة للتعديل في الثقافة المستهدفة، وإذا تم تصديقها حسب معطيات البحث.

تمت الإشـارة الى ذلك في ميـريندا، مـايو غـواداجنولي ويو- ون (1984). كان لدى المقياس الرئيس (نوفاك، بونافنتورا، ميريندا، (1979-1972) وكل الترجمات ما عدا الصينية، معلمون في صفوف يضعون تقديراتهم في نهاية كل شهر حسب ملاحظاتهم حول سلوك كل تلميذ، بالإضافة الى تدوين أدائه المدرسي في سجله الخاص. عند التخطيط لتطوير استمارة اختبار في اللغة الصينية، أقنعني إميلي مياو أنه من الأفضل أن تقوم مدرسة كل صف دراسي بملاحظة وتدوين إجابة التلميذ للبنود في نهاية كل شهر عوضاً عن الاعتماد على استرجاع الإدراك الحسي (مثال لملاحظات المدرسة: صعوبة القفز على الحبل، صعوبة التقاط الكرة، صعوبة في تذكر أشياء شاهدها). وهذا ما حدث بالفعل وبرهن على نجاحه في تطوير جاهزية الاستمارة الصينية لاختبار RIPIS.

ظهرت بعض الأحداث غير المتوقعة وغير المتنبأ بها في أثناء عملية التكييف الثقافية لعدد من الأدوات في اللغة الانكليزية إلى لغة وثقافة بلدان مختلفة لوضعها ضمن مشاريع الأبحاث عبر الثقافات بين الأعوام 1967-1995. على الرغم من الخطوات الحذرة التي تم اتخاذها، ونوقشت سابقاً، لم يتم منع حدوث بعض المشكلات التي تطلبت التصحيح. خُصص مشروع المواهب في سيسيليا الذي بدأ في 1967، في عامي الأول لاشتراكي في مشروع التبادل الثقافي بين إيطاليا والولايات المتحدة وعملي في مخبر العلوم النفسية التطبيقية، جامعة باليرمو، لتطوير مجموعة اختبارات لتعيين المواهب الكامنة لدى شباب سيسيليا.

في مرحلة تعيين المواهب في المشروع، تم جمع المعطيات باستخدام مجموعة أدوات مترجمة منها اختبار المهارة في المفردات (MOS-1). يقدم هذا الاختبار إلى الممتحن قاموس تعريف كلمة مختصر بالإضافة الى الحرف الأول للكلمة يتبعها عدد من الفراغات حسب عدد الأحرف المطلوبة لإكمال الكلمة. أحد هذه البنود في الاختبار الأصلي يقول: "نقود، خاصة نقود جاهزة، عملة أو ما يعادلها، تُدفع مباشرة

بعد الشراء" . الإجابة الصحيحة في اللغة الانكليزية هي "نقداً" في مفتاح التصحيح. في النموذج الإيطالي تم الإشارة الى حرف (ن). كلمة مكونة من أربعة أحرف تبدأ بحرف (ن)، ولكن الكلمة الإيطالية تتكون من 8 أحرف. يتوقع من طلاب المرحلة المتقدمة في إيطالية معرفة الكلمة في اللغة الإيطالية، كما يعرف الكلمة بالإنكليزية طلاب المرحلة المتقدمة في الولايات المتحدة (ميريندا، كلارك و جابكس، 1965، ميريندا، هول وباسكال. 1962، ميريندا، جابكس و كلارك، 1966).

كانت الاختبارات في مجموعة اختبار المهارة في المفردات حلزونية (تدور حول نقطة مركزية). عندما تم إقامة تحليل البند لاختبار Mos باستخدام معطيات إيطالية، كشف ذلك البند صعوبة المستوى، أي أن قيمة الاحتمال للجواب الصحيح كانت منخفضة بعكس قيمة الاحتمال العالية المعروفة في الولايات المتحدة. بدراسة لاحقة لإجابات الأفراد تم الكشف عن أن كثيراً من الطلاب الإيطاليين أجابوا بالكلمة الصحيحة كما جاءت في مفتاح التصحيح (كمبيالة) وكانت هذه مصادفة لأن الكلمة لا تتماشى مع تعريفها في الاختبار، ولكن تم تعريفها إجابة صحيحة.لهذا السبب، بإجابتين صحيحتين في البند في الاختبار الإيطالي، ارتفعت قيمة الاحتمال (p) لتتماشى مع قيمة الاحتمال في الولايات المتحدة.

قبل استخدام بطاقة الائتمان في إيطالية، كان استخدام الكمبيالة طريقة شائعة في الثقافة الإيطالية خاصة بين الطبقة العاملة. كانت الكمبيالة مماثلة للعملة الورقية وكانت تستعمل لدفع الفواتير إذا لم تتوفر العملة النقدية. ولكن، الآن في إيطاليا يجري استخدام بطاقة الائتمان؛ لذلك إذا تم استخدام Mos -1 اليوم في إيطاليا يجب تعديل مفتاح التصحيح، وتعطى درجة للإجابة الصحيحة (بطاقة ائتمان).

في المحاولة الأولى لتطوير اختبار RIPIS الى الفارسية، أثناء عملية تحضير المعطيات لتحليل عامل الأداء التي يقوم بها المدرسون في إيران، توقف الكمبيوتر

فجأة عن العمل. بعد البحث، تبين أن هناك خطأ في البند 25، " أرجع الورقة التي يوجد فيها الكثير من الكلمات الممحوة". وكانت ببساطة غير مناسبة للنموذج الذي يزود الدراسة الاستطلاعية بالمعطيات. في نقطة تفعيل جدول الارتباط، لم يكن بالإمكان حساب معامل الارتباط بين ذلك البند وبقية بنود الاختبار؛ لأنه حصل على صفر في التباين. كانت إجابة المدرسين لذلك البند "ابدأ" تعطى نقطة واحدة من درجات المقياس. بذلك، أعطى تقدير البند متوسطاً واحداً وانحرافاً معيارياً صفراً. عند تحليل السبب غير الطبيعي لتلك النتيجة مع الباحثين الإيرانيين المشاركين، تقرر أن أقلام الرصاص والممحاة لا تستخدم في المدارس التي أخذت منها النماذج. وعوضاً عن ذلك تستخدم الفرشاة والطلاء في تعليم التلاميذ الصغار الذين يبدؤون في دراسة اللغة الفارسية. على كل حال، لا تتبع كل المدارس في إيران ذات الطريقة التي تستخدم في تلك المدارس. كان ذلك مصادفة لم تكن بالحسبان. كان المقصود أن يكون ذلك النموذج متطابقاً، يمثل التلاميذ الصغار في المدارس الحكومية من مختلف الطبقات الاجتماعية/ الاقتصادية. في منطقة فقيرة أو في مدارس خاصة من مناطق طبقة غنية (ميريندا. 1990b).

تم إدراك المشكلات التي واجهها باحثو تطوير الاختبار المكيف بسرعة وجرى إصلاحها. وتوصف الآن لتوضح ما يمكن أو ما يحدث حتى في حال اتخاذ الخطوات الضرورية عند تكييف الاختبار للاستخدام في ثقافة جديدة. عندما تكون الترجمة هي الخطوة الوحيدة المتخذة في نقل الأداة من ثقافة الى أخرى، بالرغم من دقة الترجمة، يمكن حدوث بعض المشكلات دون أن يتم اكتشافها. لكي يتم تكييف الاختبار بشكل سليم يجب جمع معطيات حقيقية، وتحليلها وشرحها.

تجربة حقيقية أخرى

قبل البدء في ذكر الأحداث الجدية والمؤسفة التي يتم حدوثها بالممارسة الخاطئة عند نقل أداة القياس من ثقافة الى أخرى باستخدام الترجمة فقط، التي

غالباً ما تكون خاطئة، نقوم بوصف هذه الأحداث: في أوائل السبعينيات كنت أعمل في المجلس الاستشاري في بتراليا سوبرانا/ سيسيليا. كان الأعضاء الآخرون في المجلس سكارفيا آندرسن، ساموئيل ميسيك. ميريام غولدنبرغ، مارغريت ميد وتلميذتها التي كانت تُعد رسالة الدكتوراة جوزفين دانا.

كان كاهن البلدة، كالجرو لابلاسا، قد أسس المدرسة في تلك البلدة الجميلة. بعمله الدائب، استطاع الأب لابلاسا الحصول على تمويل خارجي ليس لتأمين تعليم مجاني فقط، ولكن لتأمين إقامة مجانية للأطفال الموهوبين ولتوسيع منشآت المدرسة سنوياً. كان الأب لابلاسا يمضي أكبر قسم من العام بزيارة المدارس لإيجاد واختيار الأطفال الموهوبين للصفوف في المستقبل، ولطلب الإذن بالحديث مع الطالب الذي تعتبره إدارة المدرسة الأفضل في المدرسة. بعد ذلك يتحدث الأب لابلاسا مع الطالب، وإذا كان ذلك الطالب مهتماً بالالتحاق بالمدرسة، عندئذ يتحدث مع الأهل لإقناعهم بالسماح لولدهم في أن يغادر البلدة/ المدينة ويلتحق بمدرسة الموهوبين.

كان من عادتي أن أمضي الصيف في سيسيليا أعمل على مشروع بحثي الخاص في جامعة بالميرد. وكان الأب لابلاسا يتطلع دائماً الى تمضية بعض الوقت في التشاور معي. ذات صيف، في لقائنا الأول ذلك العام، أخذني بعيداً عن زوجتي ووضعني مع كاهن آخر في سيارة ثانية عند ذهابنا الى الغداء. كان الأب لابلاسا منزعجاً جداً وأراد أن أتناقش مع الكاهن في موضوع كان قد رواه للأب لابلاسا. كان الكاهن الذي كان قد حضر دورة في إقامة ووضع الدرجات وتفسيرها لاختبار النسخة الإيطالية لمقياس ذكاء فكسلر للأطفال، قد أجرى الاختبار لكل الطلاب المسجلين في المدرسة. ما عدا طالب "موهوب جداً" كان قد حصل على مجموع 108 في معامل الذكاء، أحرز الطلاب الآخرون درجات ضمن الحدود العادية للدرجات (70-90). من غير الضروري القول إنني أمضيت أكثر الوقت في الطريق الى المطعم أشرح للكاهنين خطأ اختبار أطفال إيطاليين باستخدام اختبار جرى تحضيره،

وقياسه ووضع معاييره في الولايات المتحدة وتمت ترجمته فقط إلى اللغة الإيطالية. ما تم ترجمته لم يكن بنود الاختبار فقط، ولكن إجراءات التصحيح وكيفية تعيين درجات المقياس ومعامل الذكاء، والوصف الكمي لمدى الذكاء حسب المعايير في الولايات المتحدة. بذلك، تكون النتائج حسب ما توقعته.

الأحداث الجدية التي من الأسف حدوثها

إن المشكلات التي جرى حدوثها خلال الوقت الطويل للبحث في تكييف أدوات الاختبار عبر الثقافات، كانت جدية لكنها لم تكن مدمرة. بالرغم من أنها كانت سبب التأخير في تقدم تكييف ناجح ومضاعفة الكلفة من حيث الوقت والمادة والجهد. ولكن جرى إيجاد الحل لها في النهاية. على كل، بعض المشكلات الأخرى التي لم يتم معالجتها فوراً أحدثت أكثر من تأخير. كانت الإحداثيات الأساسية (أ) إحداث ترجمة للأداة من اللغة الأصلية الى أخرى، و (ب) الافتراض أن الأداة الأصلية جيدة. كانت النتائج مدمرة لكل الذين اشتركوا في تلك العملية. دعوني أشرح لماذا.

في أوائل عام 1928 كان ليندرو آلميدا، أستاذاً مساعداً في قسم العلوم النفسية والعلوم التربوية في جامعة بورتو/ البرتغال، طالب دكتوراة في جامعة لوفان/ بلجيكا. كان أستاذه والمشرف على رسالته جورج موريس مؤلف كتاب " مجموعة اختبار موريس للاستدلال المنطقي الفارق". حضر ألميدا، الذي أصبح فيما بعد عالم نفسي رائد في البرتغال ومؤلف الاختبار المكيف البرتغالي الميدا، (1988)، إلى الولايات المتحدة لطلب الاستشارة حول الاختبارات النفسية والقياسات النفسية. اختار ثلاثة من الباحثين لاستشارتهم: روبرت روسنثال (هارفارد)، روبرت ستينبرغر (ييل) وبيتر ميريندا (جامعة رود آيلند). جاء أولاً الى كامبردج ولكنه علم لاحقاً أن روسنثال كان في إجازة بحث علمي ولم يكن موجوداً. عندئذ جاء إلي قبل ذهابه إلى نيو هايفن للقاء ستينبرغر. بهذا، بدأت علاقة مهنية طويلة استمرت حتى اليوم.

في عام 1984 تقاعدت مبكراً من جامعة رود آيلند. كان السبب الرئيس لذلك التقاعد هو قبولي منصباً قصير الأجل في جامعة في الخارج لتدريس القياسات السيكولوجية لأعضاء كليات علم النفس الذين كان عليهم القيام بأبحاث حقيقية للحفاظ على مناصبهم أو التقدم في رتبتهم. كان عملي الرئيس بين الأعوام 1984-1993 في جامعات ليشبونة، بورتو، مينهو في البرتغال. خلال تلك المدة (كما في كل بلدان البحر المتوسط) كما في إيطاليا واليونان، كانت الجامعات تمنح شهادة واحدة فقط، شهادة البكالوريوس. كان الكادر التدريسي في المعاهد أو أقسام علم النفس مكوناً من أستاذ وحيد فقط أو عدد قليل من الأساتذة. كانت مسؤولية تدريس الطلاب تقع بشكل أساسي على عاتق " المساعدين" الحاصلين على شهادة البكالوريوس في علم النفس. كان في البرتغال مستويات للمساعدين، "أ" أو "ب" (أو كما يسمونه الآن الأستاذ المساعد أو الأستاذ المشارك). كان البحث الموافق عليه شرط أساسي للتقدم من المرتبة "أ" إلى المرتبة "ب". وكان يجري استفتاء عن المساعدين الذين لم ينجحوا في كليات الجامعة ليصبحوا أساتذة في المدارس في أحسن الأحوال؛ لذلك كانت الأبحاث المقبولة التي يمكن الدفاع عنها أمام اللجنة الفاحصة تأخذ الأولوية في الإنجاز من المساعدين في مرتبة "أ".

كانت حلقات التدريس قصيرة المدى التي كنت أدرسها يتضمن أغلبها تدريس الإحصائيات المتعددة التباين، منهجية البحث، تحليل العوامل، طرق القياس السيكولوجي، وحلقات معالجة الكمبيوتر. كانت تلك الحلقات شبيهة بالتي كنت أدرسها في الولايات المتحدة. بالإضافة الى تدريس تلك الحلقات في البرتغال، كنت ألتقي بشكل دوري مع المساعدين لمراجعة اقتراحات بحثهم وإعطائهم النصح. وكان لي بعض التحفظات على بعض الأبحاث التي كانوا يقومون بها. كان المساعدون يقومون باختبار أداة تقويم في نطاق علم النفس، التي يهتمون بها في إقامة البحث ثم يجرون ترجمتها الى البرتغالية.

كانت إجراءات جمع النماذج في البحث والطرق الإحصائية المستخدمة في تحليل المعطيات المبنية على تلك الأدوات المترجمة (بشكل أساسي من اللغة الإنكليزية الى البرتغالية) صحيحة بشكل عام ودقيقة. على حال، في بعض الحالات، بعد عدة شهور من العمل الجاد والتكاليف الباهظة تحدث الكارثة. عندما كنت أعود إلى البرتغال في العام التالي أو بعده كنت أجد المساعدين ينتظرون عودتي بفارغ الصبر. كان السبب الأساسي لذلك أن نتائج تحليل معطيات ذلك البحث لم يكن لها معنى، لا للباحثين ولا للمشرفين على البحث. وأكثر من ذلك كانت هناك مشكلات متعلقة بإثبات عوامل بنية النسخة المترجمة للأداة، وغالباً عدم وجود معنى للبنية.

فيما يلي نناقش نموذجين من تلك التجارب التي كانت مخيبة للآمال ومحبطة لي ومؤلمة للباحث.

1- أكملت مساعدة في علم النفس في جامعة في البرتغال كانت تسعى للحصول على شهادة الدكتوراة من جامعة في فرنسا، قسماً أساسياً في بحث رسالتها للدكتوراة. كانت قد أنهت كتابة عدة فصول في رسالتها، وكانت تواجه عدة صعوبات في شرح معطيات كثيرة كانت قد عالجتها خلال وجودها في فرنسا كطالبة دكتوراة متفرغة بشكل كامل. كانت تنتظر وصولي لأبدأ مهامي التدريسية في كليتها بفارغ الصبر مع معطياتها الغامضة. لخيبة أملي ولكدرها، لم نأخذ وقتاً طويلاً لاكتشاف أساس المشكلة.

كانت أدوات التقويم التي جرى تقديمها بالطرق التمهيدية لتحليل العوامل خاطئة، لأنه لم يجر تكييفها ثقافياً بشكل جيد وتام. كان التحليل يتضمن جداول منفردة، وذلك أدى الى استخلاص جذور سلبية مستترة. لإقامة تحليل عوامل صحيح لمصفوفة ارتباط، يجب عليه أن يمتلك خاصية كونه "gramian" إن مصفوفة عامل ارتباط هي تلك التي تكون متماثلة والتي تكون جذورها الكامنة المستخلصة كلها إيجابية أو قريبة من الصفر أو صفراً. ويطلق على مصفوفة الارتباط تلك

"إيجابية شبه محددة" أو "إيجابية محددة". وكل مصفوفات الارتباط الأخرى كانت تعرف كمصفوفات ارتباط مفردة ولا يجب تحليلها بأية طريقة من تحليل العوامل. إذا تعرضت مصفوفات ارتباط مفردة الى تحليل عوامل فإنها ستعطي قيمة مستترة ولن يفي التوازن الأساسي بشروط تحليل العوامل (ميريندا، 1997، تاتسوكا، 1971) لذلك فإن نتائج تحليل العامل تكشف عن عدم وجود أية بنية للأدوات. إن حمولة العامل لأكثر البنود كانت حوالي 20s أو أقل. والكثير منها كانت قريباً من الصفر. جعلت تلك النتائج من المستحيل أن أعطيها أية نصيحة عن كيفية تعديل الأدوات وعن إعادة معايرة الدرجات.

2- كانت التجربة تخص إحدى المساعدات التي كانت سابقاً إحدى تلميذاتي، عندما ساعدتها سابقاً بإكمال مشروع بحث اختصاصي بنجاح. وذلك ساعدها في الحصول على مرتبة أعلى. كانت قد التحقت مرة ثانية بأحد صفوفي، وأنهت مشروع بحث آخر وكانت تسعى الى ترقية أخرى. في مرحلة معالجة المعطيات في بحثها، كانت قادرة على شرح الكثير من المعطيات التي سبق تحليلها. ولكنها كانت مرتبكة من نتائج استبيان لثقافة من خارج البرتغال، والذي كان أساسياً في الفرضية الرئيسة للدراسة. كانت تلك المعطيات متعلقة بالإجابات التي تم جمعها بإقامة عملية مسح على أمهات تلاميذ مدارس فعلية.

عندما عدلت الاستبيان لتكييفه حسب الثقافة البرتغالية، اتبعت إرشاداتي السابقة في كيفية ترجمة البند بشكل صحيح وكيفية تغيير بنية إجابة البند من درجات ثنائية القيمة الى فاصل مقياس تقدير بفواصل 5 أو 7 نقاط. لسوء الحظ كانت النتائج غير قابلة للتفسير. خامرني الشك في أن معاملات الارتباط المتبادلة لم تكن صحيحة للاستخدام بتحليل العوامل لعدم حيازتها على المواصفات المطلوبة. كانت، كما في التجربة الأولى، ليست معاملات ارتباط مفردة (ميريندا، 1997، تاتسوكا، 1971). عند معاينة نتائج جداول القيمة المستترة، قدمت كل الجداول قيمة

مستترة سلبية. ماذا سبب تشويه في مصفوفات معاملات الارتباط الجواب موجود في معاينة توزع مقياس إجابة البند ذي النقاط الخمسة أو السبعة. كان بعضهما غير مترابط وكان الآخر غير كامل أو كان مشوهاً بشكل كبير. لم تكن البنود الموجودة في الاستبيان قد جرى تكييفها ثقافياً. أو من الممكن أن الاستبيان الأصلي لم يكن صحيحاً حسب القياس السيكولوجي. لم يكن هناك أي خيار للباحثة إلا التضحية بسنتين من العمل المكثف في محاولة لتكييف الاستبيان والبدء من جديد.

*****

المراجع

Almeida, L. S. (1988). *Oraciocinio differencial dos jovens* [A reasoning differential test for juveniles]. Porto, Portugal: Instituto Nacional de Investigacas Cientifica.

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* Thousand Oaks, CA: Sage.

Berry, J. W. (1997). Immigration, acculturation, and adaptation. *Applied Psychology: An International Review, 40,* 5–68.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods.* New York: Wiley.

Calvi, G. (1970). *Il Test G.* Firenze, Italy: Organizzazioni Speciali.

Clarke, W. V. (1956). The construction of an industrial selection personality test. *Journal of Psychology, 41,* 379–394.

Clarke, W. V. (1960). *Examiner's manual for the measurement of skill: A battery of practical placement tests.* Providence, RI: Walter V. Clarke Associates, Inc.

Cronbach, L. J., & Drenth, P. J. D. (Eds.). (1972). *Mental test and cultural adaptations.* The Hague, Netherlands: Mouton.

D'Amico, G., Merenda, P., & Sparacino, R. R. (1982). *Rhode Island Pupil Identification Scale (R.I.P.I.S.): Primo adattamento Italiano.* Palermo, Sicily: Stass.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycroft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A. (1964). *The American high school student* (Cooperative Research Project No. 635). Pittsburgh: Project Talent Office, University of Pittsburgh.

Geisinger, K. F. (1994). Cross-cultural normative assessment, translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Bulletin, 106,* 304–312.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings.* Yonkers, NY: World Book.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests [special issue]. *Language Testing, 20*(2), 127–240.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1,* 1–13.

Healy, W., & Fernald, G. M. (1911). Tests for practical mental classification. *Psychological Monographs, 13*(Whole No. 54).

Knox, H. A. (1914). A scale based on the work at Ellis Island. *Journal of the American Medical Association, 62,* 741–747.

Kohs, S. C. (1920). The block-design tests. *Journal of Experimental Psychology, 3,* 357–376.

Lecky, P. (1945). *Self-consistency: A theory of personality.* New York: Island Press.
Marston, W. M. (1928). *Emotions of normal people.* New York: Harcourt.
Marston, W. M., King, C. D., & Marston, E. H. (1931). *Integrative psychology.* New York: Harcourt.
Merenda, P. F. (1990a). Present and future issues in psychological testing in the United States. *Evaluacion Psicologia/Psychological Assessment, 6,* 3–31.
Merenda, P. F. (1990b). The Rhode Island Pupil Identification Scale (RIPIS) in cross-cultural perspective. In L. L. Adler (Ed.), *Cross-cultural research at issue.* New York: Academic Press.
Merenda, P. F. (1993). Cross-cultural current and future issues in psychological testing. *International Journal of Group Tensions, 23,* 115–132.
Merenda, P. F. (1994). Cross-cultural testing: borrowing from one culture and applying it to another. In L. L. Adler & E. P. Gielen (Eds.), *Cross-cultural topics in psychology.* Westport, CT: Praeger
Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development, 30,* 156–164.
Merenda, P. F., Clarke, W. V., & Jacobsen, G. (1965). Relative predictive validities of the MOS and DAT batteries for junior high school students. *Psychological Reports, 16,* 151–154.
Merenda, P. F., & DiLeonardo, C. (1992). The American adaptation of the Bedini Test of Distributive Attention. In A. L. Comunian & U. P. Gielen (Eds.), *Advancing psychology and its applications: International perspectives.* Milan, Italy: Franco Angeli.
Merenda, P. F., Hall, C. E., Clarke, W. V., & Pascale, A. (1962). Relative predictive effeciency of the DAT and a short battery of tests. *Psychological Reports, 11,* 71–81.
Merenda, P. F., Jacobsen, G., & Clarke, W. V. (1966). Cross-validities of the MOS and DAT batteries. *Psychological Reports, 19,* 341–342.
Merenda, P. F., Maio, E., Guadagnoli, E., & Yu-Wen, H. (1984). *The Chinese form of the Rhode Island Pupil Identification Scale: Standardization and validation.* Providence, RI: AVA Publications.
Merenda, P. F., & Migliorino, G. (1974). Student information blank: Comparison between Sicilian and American responses. *Annali della facolta di Economia e Commercio, Dell' Universita' "di Palermo, 28,* 385–403.
Migliorino, G. (1947). Ricerca sulla strattura psicologica del reattivo di Bedini. *Rivista di Psicologia, 43,* 154–171.
Novack, H. S., Bonaventura, E., & Merenda, P. F. (1979). *Manual to accompany Rhode Island Pupil Identification Scale: A behavior observation identification scale: A behavior observation scale for the early detection of children with learning problems.* Providence, RI, Author. (Origianl work published 1972)
Olmedo, E. L. (1979). Acculturation: A psychometric perspective. *American Psychologist, 34,* 1061–1070.
Pintner, R., & Paterson, D. G. (1917). *A scale of performance tests.* New York: Appleton.
Poortinga, Y. (1995). Use of tests across cultures. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 187–206). Boston: Kluwer.
Porteus, S. D. (1915). Mental tests for the feeble-minded: A new series. *Journal of Psycho-Asthenics, 19,* 200–213.

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence.* London: Lewis.

Rulon, P. J. (1953). *A semantic test of intelligence. Proceedings of 1952 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Sperber, A. D., Develles, R. F., & Boehlecke, B. (1994). Cross-cultural translation. *Journal of Cross-Cultural Psychology, 25,* 501–524.

Sprini, G., Cardica, M., & Gangemi, A. (1992). *The Italian form of the RIPIS scale: A format modifying proposal.* Palermo, Sicily, Italy: University of Palermo.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., & Clausen, J. A. (1950). *Measurement and prediction.* Princeton, NJ: Educational Testing Service.

Tatsuoka, M. M. (1971). *Multivariate analysis: A technique for educational and psychological research.* New York: Wiley.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Woodworth, R. S. (1910). Race differences in mental tests. *Science N. S., 31,* 171–176.

14

تقييم التقاطع الثقافي للحالات الوجدانية والسمات الشخصية

تشارلز د. سبيلبرغر، مانوليت س. موسكوسو
وتوماس م. برونر
جامعة جنوب فلوريدا

في المقطع التمهيدي لهذا الكتاب، يستعرض هامبلتون مجموعة واسعة من الموضوعات العامة المتعلقة بتكييف التقاطع الثقافي (عبر الثقافات) لمقاييس الإنجاز والأهلية والشخصية. وقد تم شرح مصادر مهمة للأخطاء المصادفة في تكييف الاختبارات، وقدمت توجيهات لتقليل الخطأ وتحسين صدق الاختبار (هامبلتون، 1994، هامبلتون وباتسولا، 1999، فان دي فيفر وهامبلتون، 1996). وقد تم بالتفصيل بحث ثلاث فئات واسعة من الموضوعات والمشكلات التي تمت مواجهتها في تكييف الاختبار. في هذه المنشورات (أ) الاختلافات اللغوية والثقافية، (ب) العوامل التي تؤثر على تفسير نتائج الاختبار. وقد تم عرض ومناقشة توجيهات عملية طورت حديثاً من قبل لجنة الاختبار العالمية (ITC) لترجمة وتكييف الاختبارات النفسية والتربوية.

إن توجيهات وتكييف الاختبار ITC تقدم توصيات ممتازة لطرق وإجراءات تكييف التقاطع الثقافي للاختبارات النفسية والتربوية. إن اتباع هذه التوجيهات، يعد شيئاً أساسياً في مقاييس التكييف للإنجاز والأهلية وذلك لتسهيل المقارنة لأداء الطلاب من مختلف اللغات والثقافات. بينما تعتبر توجيهات وتكييف الاختبار ITC

قابلة للتطبيق لتكييف كل أنواع الاختبارات النفسية، تختلف الميزات الشخصية والحالات الوجدانية عن الكفاءات والقدرات (أنستازي، 1988). إن الحالات العاطفية والسلوكيات التي تشكل الميزات الشخصية هي أكثر شخصية، وأقل وضوحاً بالتعريف من الكفاءات والقدرات والإنجاز. علاوة على ذلك، كما لاحظت أنستازي " بشكل يفوق حتى اختبارات القدرة، يمكن التوقع أن اختبارات الشخصية ستظهر اختلافات كبيرة في الثقافة الثانوية تماماً كالاختلافات الثقافية" . (ص532).

تعزى أيضاً الاختلافات في تفسير تعليمات الاختبار إلى المشكلات في تكييف التقاطع الثقافي لمقاييس العواطف والشخصية. مثلاً، لاحظت مارسيلا ولونغ (1995) أن الأشخاص من الثقافات غير الغربية قد لا يرتاحون لإعطاء أجوبة صح -أم- خطأ لأسئلة مينوسوتا المتعلقة باختبار الشخصية المتعدد الأوجه (MMPI) لأن الأشخاص من الثقافات الاشتراكية يضعون بشكل نموذجي تأكيداً أكبر على عوامل الموقف الذي يؤثر على مشاعرهم وتصرفاتهم. لتوضيح هذه النقطة، قام كل من مارسيلا ولونغ (1995) باقتباس إجابة فيلبيني لـ MMPI: "سيدي، هذا يبدو صحيحاً أحياناً وخاطئاً أحياناً أخرى. لا أستطيع إخبارك هل هو صحيح أم خطأ كل الوقت" (ص208).

إن تكافؤ التركيب يعد مطلباً أساسياً في تكييف تقاطع الثقافات لكل أنواع الاختبارات. و"يجب أن تتم العناية باختيار الحالات، والمفردات والتعبيرات التي سوف تتكيف بسهولة عبر المجموعات اللغوية والثقافية" (هامبلتون، انظر إلى المقطع 1 من هذا الكتاب). في تكييف مقاييس الشخصية والمشاعر، يجب أن يبذل اهتمام خاص إلى حالة السمات المميزة (أنستازي، 1988، كاتل، وسبيلبرغر، 1960، كوهين، سويردليك وسميث، 1992، لونر، 1990، وسبيلبرغر، 1966b). ولتحديدكثافة الاسئلة (أنستازي، 1988، وسبيلبرغر، غررسوش ولاشين، 1970) في تقييم الاختلافات الفردة بالسمات الشخصية، يجب أيضاً أن يتم تقييم التكرار المتعلق بحدوث الحالات النفسية (وسبيلبرغر، 1983، 1988).

إن عدم تكافؤ التراكيب في اللغات والثقافات المختلفة هو المصدر الأكثر خطورة للخطأ في مقاييس تكييف الشخصية والمشاعر، كما لاحظ هامبلتون في المقطع الأول من هذا الكتاب. إن تكافؤ التقاطع الثقافي شيء صعب، خاصة للحصول على مقاييس الشخصية. لأن هناك، حتى الآن، توافقاً قليلاً نسبياً يتعلق بمعايير تحديد الأبعاد الشخصية الأساسية (كوهي وآل، 1992، كرونباش، 1960، هال وليندي، 1970). مثلاً، هناك ترابط محدود فقط بين مقاييس المتزامنات السريرية الذي يرتكز عليه مقياس MMPI وأبعاد الشخصية المقدرة من قبل الـ MMPI. إن إدراك هذه العيوب قد دفع تطوير معايير المضمون لـ MMPI لتقدير القلق والحزن والإحباط والغضب والمتغيرات الشخصية الأخرى المشابهة (بوتشر، غراهام، ويليامز، وبن- بوراث، 1989).

خلال الـ20 عاماً المنصرمة، تلقت الأبعاد المدعوة بأبعاد الشخصية "(الخمسة الكبيرة" قبولاً كبيراً كتراكيب شخصية أساسية (مثال، ديغمان، 1990، غولدبرغ، 1981، جون، 1990) إلا أن العصابية، وهي واحدة من الأبعاد الشخصية الخمسة الكبيرة المقدرة من قبل مقياس الشخصية NEO المستخدم بكثرة PI-R: NEO: كوستا وماكر1992 وأيضاً من قبل استفتاء الشخصية آيسينغ لـ : EPQ آيسينغ وآيسينغ، (1987) تعد تركيباً متغايراً كثير التعقيد مؤلفاً من عدد أكبر من الأبعاد الأساسية. انعكس هذا التعقيد في تعريف القلق، والغضب - العدائي، والإحباط، وهي ثلاثة من المظاهر الأساسية للعصاب، المقدرة من قبل المقاييس الثانوية لـ NE OPI كوستا ومكريت، 1992) ويمكن اعتبار هذه المظاهر كأبعاد أساسية للأعراض المتزامنة بالعصاب.

يركز هذا المقطع على المشكلات الفريدة نسبياً التي تتم مواجهتها في تكييفات التقاطع الثقافي لقياس المشاعر والشخصية. نقوم أولاً باختبار المشاعر والشخصية كتراكيب نفسية. ثم ندرس تكافؤ التقاطع الثقافي لمفاهيم المشاعر والشخصية في سياق متجدد، وكيف تؤثر الاختلافات الثقافية على معاني الكلمات المستخدمة

لوصف هذه التراكيب. إن الحاجة الكبيرة لأخذ حالة السمات المختلفة بعين الاعتبار في تكييف مقاييس الحالات العاطفية والسمات الشخصية تحلل بعد ذلك. وتتم مناقشة أمثلة لثقافة معينة لتكييف مقاييس القلق من الإنكليزية إلى اللغات الأخرى. أيضاً، يتم اختبار مؤثرات اللغة والثقافة في تكييف مقاييس الخبرة، والتعبير والسيطرة على الغضب في ثقافات ناطقة باللغة الإسبانية.

قياس السمات الشخصية والحالات الوجدانية:

تبعاً لهال وليندزي (1970)، "لا يوجد تعريف جوهري للشخصية يمكن تطبيقه بأية أغلبية" (ص9). تتراوح تعاريف الشخصية من الاعتبارات الشاملة للسلوك بكل تفاصيله المعقدة، وحتى المواصفات المحددة للسمات الشخصية (اناستاسي، 1988، غوثري ولونر، 1986) أكدت اناستاسي على أهمية تعريف الشخصية من حيث المفاهيم التي تصنف الفئات إلى أي سلوك يجب أن تصنف في حال تم قياسها بدقة. بالتوافق مع برهان أناستاسي حول سمات الشخصية الأساسية، عرف كوهين وآخرون الشخصية بكونها " مجموعة فريدة من الحالات والسمات النفسية لفرد ما" (ص401). يعد الغضب والقلق والفضول أمثلة للحالات والسمات الهادفة التي ترتبط بشكل استثنائي بالشخصية (وسبيلبرغر، ريهيسير، وسيدمتن، 1995).

تم تبسيط تكافؤ التقاطع الثقافي للقلق والغضب كحالات وجدانية وسمات شخصية عبر حقيقة أن هذه المشاعر الأساسية برهنت على أنها نتاج عالمي للتطور. في كتابه الكلاسيكي، تعابير المشاعر عند الإنسان والحيوان، استنتج داروين (1872، 1965). كما أكد آخرون (إكمان، 1973، إزارد، 1977، تومكينز، 1962) أن الخوف والغضب مشاعر مكثفة يمكن أن تعرف من تعابير الوجه، ليس فقط لدى البشر، ولكن لدى عدة أنواع من الحيوانات.

بالربط مع هذه الاكتشافات البحثية، لاحظ ديمبيرغ (1994، 1998) أن ردات الفعل المميزة المتعلقة بالوجه ظهرت بعد تعرض قصير جداً لمنبهات خاصة بالخوف

والغضب، كالأفاعي والوجوه الغاضبة. وهذا يشير إلى أن استقبال منبهات مهددة يمكن أن يعمل على تحريض مشاعر خاصة على الفور.

لقد قدم بلوتشيك (1984) نظرية "التطور النفسي" التي تحدد المشاعر والحالات المعقدة التي يمكن استنتاجها من التقارير الذاتية، التغيرات النفسية والأشكال المتعددة للسلوك الذي يمكن فهمه بشكل الأفضل في سياق متطور. تأييداً لمنظور داروين المتعلق بعلم دراسة سلوك الحيوان، أشار بلوتشك (1984) إلى الدور التكيفي للمشاعر في تحفيز ما وصفه كانون (1963) بسلوك ردات الفعل "قاتل - أو- أهرب"، إلى حالات الطوارئ البيئية التي تزيد فرص نجاة الكائن الحي. ولكن، كما لاحظ بلوتشك أن وصف المشاعر المترافقة مع ردات الفعل السلوكية، سوف تعتمد على خبرة الشخص في لغة معينة.

إن الكلمات المستخدمة في لغات متعددة لوصف الحالات الوجدانية والسمات الشخصية لديها، بشكل عام، مجموعة واسعة من الدلالات (روغلر، 1999، ويرزبيكا، 1994). حتى ضمن لغة معينة، قد يكون في الكلمة نفسها اختلاف في المعاني في الثقافات البديلة (اناستاسي، 1988). لذا، فإن الفروقات في، وضمن الثقافات، في معاني الكلمات التي اعتادت أن تصف الحالات الوجدانية والسمات الشخصية، تعد جدلية بشكل خاص في تكييف التقاطع الثقافي لمقاييس هذه التراكيب (روغلر، 1999) وفيما يلي أمثلة عن فروقات الثقافة البديلة في معاني كلمات إسبانية:

- في دول بحر الكاريبي تعني كلمة "guagua"حافلة (باص)، ولكن هذه الكلمة نفسها تدل على رضيع أو طفل، في تشيلي وكولومبيا وبيرو.
- تعني "verraco" خنزير في كوبا، ولكن في كولومبيا، لها دلالات لشخص جلف.

- في كوبا، تشير bicho إلى حشرة، ولكنها تصف قضيب الرجل في بورتوريكو.
- في إسبانيا، كلمة coger لها معنى غير مؤذ يعبر عن الأخذ أو الإمساك، ولكنها تعني ممارسة الجنس في المكسيك وفنزويلا.

تشير هذه النماذج بوضوح إلى أن التكييف الناجح لمقاييس التقرير الذاتي المتعلقة بالحالات الوجدانية والسمات الشخصية تتطلب الاختيار الدقيق لمجموعة مفاتيح الكلمات (أو التعابير) التي لها، جوهرياً المعنى نفسه في كل من لغة الأصل (المصدر) واللغة الثانية (الهدف). ولكن تأكيد التمثيل الدقيق للمفاهيم النفسية المقدرة، غالباً ما يكون شيئاً صعباً لأن اللغات تختلف في دلالات الكلمات المستخدمة لوصف المشاعر والإدراكات المصاحبة لحالات وجدانية مختلفة وسمات شخصية. علاوة على ذلك، كما لاحظ ويرزبيكا (1994). إن مجموعة التعابير الوجدانية المتوفرة في أية لغة مفترضة هي شيء فريد ويعكس منظوراً فريداً للثقافة على أساليب الناس الوجدانية" . (ص135).

لا يمكن لمقاييس التقرير الذاتي للقلق والمشاعر الأخرى أن تترجم وتترجم إرجاعياً ببساطة. ولكنها يجب أن تكيف من أجل أبحاث التقاطع اللغوي. تستخدم عملية "الترجمة الارجاعية" بشكل تقليدي لتسهيل تكييف الاختبارات التربوية والنفسية من لغة إلى أخرى (بريسلين، 1970، 1986) في الترجمة الإرجاعية لأسئلة الاختبار من اللغة الهدف إلى اللغة الأصلية، يتم تأكيد الترجمة الحرفية للكلمات. ولكن الترجمة الارجاعية لسؤال معياري أصلي غالباً ما يكون أقل ملاءمة من تركيب سؤال جديد يرتكز على التكافؤ المفهومي لتقاطع الثقافات للحالة الوجدانية أو الأبعاد الشخصية التي يجري قياسها (سبيلبرغ وديلز، غورو، 1983) وهذا يصح بشكل خاص في تكييف تعابير المصطلحات.

زعم ليكومب وأونر (1976) أن ترجمة مفاتيح الكلمات وتعابير المصطلحات هو شيء صعب بشكل خاص، وقد يتطلب استشارات متكررة لخبراء في اللغة. انطلاقا

من وجهة نظر حرفية أو دقيقة، فلقد أوصوا أن يتم تصنيف الأسئلة إلى ثلاثة فئات. (أ) أسئلة مع مفاتيح كلمات ترجمتها تلائم بشكل قريب معنى الكلمة في لغة المصدر. (ب) أسئلة مع مفاتيح كلمات من الصعب إيجاد أسئلة مطابقة لها في لغة الهدف، و(د) أسئلة ذات صيغة لغوية لا يمكن ترجمتها من لغة المصدر إلى لغة الهدف دون تغيير التركيب النحوي. قد يُستلزم عدد من دورات الترجمة والترجمة الإرجاعية، قبل أن يمكن تطوير تكييف ملائم للنوع السابق من السؤال (سبيلبرغ وشارما، 1976).

في مقاييس تكييف الحالات العاطفية والسمات الشخصية، قد يحوي مفتاح كلمة لسؤال ما في لغة المصدر عدة ترجمات مختلفة، مقبولة بشكل متساو في لغة الهدف. قد يتم أيضاً تمثيل كلمات مفاتيح مختلفة في سؤالين أو أكثر في لغة المصدر، وذلك بواسطة كلمة واحدة في لغة الهدف. حين لا تكون الترجمة الحرفية لسؤال ما في الاختبار ممكنة، فمن المهم أن نستبقي المعنى الأساسي للسؤال الأصلي باختيار مرادف لمفتاح الكلمة التي تعكس معناها الأساسي في لغة الهدف.

يجب أن يوجه اهتمام خاص لترجمة الدلالات العاطفية للمصطلح في أثناء تكييف المصطلحات التعبيرية، عوضاً عن ترجمة المعنى الحرفي للكلمات المفردة (غولري ولونر، 1986). إن تحديد مصطلحات تعبيرية قابلة للمقارنة مع اللغة التي يتم فيها ترجمة مقياس ما، هو شيء مرجح أكثر من الترجمة الحرفية للمصطلح الأصلي. بناءً على هذا، فإن تكافؤ التقاطع الثقافي للمفاهيم النظرية التي يتم قياسها في ترجمة وتكييف المصطلحات يعد شيئاً جوهرياً. بافتراض الصعوبات التي من المرجح أن تصادف في ترجمة مفاتيح الكلمات والمصطلحات التعبيرية، فإن مجموعة أكبر من الأسئلة عندها، سوف يُستلزم في النهاية أن يتم تركيبها، من أجل إدراك المعنى الكامل للتركيب الذي يتم قياسه. يمكن بعد ذلك، استخدام الإجراءات الإحصائية لتحديد الأسئلة التي تملك ترابطاً داخلياً حسب المقاييس في التركيب المحدد.

قياس حالة وسمة القلق:

بالرغم من أن الاهتمام المعاصر بظاهرة القلق لديه جذور تاريخية في الأفكار الفلسفية واللاهوتية لباسكال وكيرغارد (أيار، 1977) فقد كان فرويد (1924، 1936) أول من حاول تفسير معنى القلق ضمن سياق النظرية النفسية. لقد اعتبر أن القلق هو "شيء محسوس". هو حالة أو وضع مؤثر مؤلم. تبعاً لفرويد (1924)، فإن هذه الحالة كما لوحظت لدى المصابين بالقلق العصبي، تم وصفها بكل ما تشمل كلمة التوتر العصبي، "nervousness" الذي يتضمن هاجساً أو توقعات مقلقة، وظاهرة التفريغ العصبي.

لا يمكن تمييز القلق عن الحالات "الوجدانية" المؤثرة المؤلمة الأخرى، كالغضب والحزن، بسبب التركيبة الفريدة لخواصه الظاهرية والنفسية. وهذا يعطي القلق "ميزة خاصة من عدم الارتياح" بحيث يبدو، بالرغم من صعوبة وصفه، "يملك سمة خاصة بذاته" (فرويد، 1936، ص69). تم تأكيد الخواص الظاهرية والشخصية للقلق -مشاعر لهاجس التوقع أو الجزع- من قبل فرويد، خصوصاً في كتاباته الأخيرة. بينما كانت السلوكيات النفسية لظاهرة التفريغ (العصبي) -على الرغم من اعتبارها جزءاً أساسياً من ظاهرة القلق لمساهمتها المهمة بالشعور بعدم الارتياح- ذات اهتمام نظري ضئيل من قبله. اهتم فرويد بشكل رئيس بتحديد مصادر الدوافع التي تثير ردات فعل القلق، أكثر من تحليل خواص حالة كهذه. وكان يأمل أن يكتشف، بخبرته السابقة مع مرضاه، "العنصر التاريخي الذي يربط عناصر القلق العصبية الناقلة والمفرغة، بشدة مع بعضها" (1936، ص70).

تم التقصي عن القلق في دراسات كثيرة، تم فيها اختيار المشاركين الذين افترض أنهم يتباينون في الدافع أو درجة الحافز (سبنس، 1958) على أساس الحد الأقصى من درجاتهم في الاستبيانات، مثل بيان مقياس القلق لتايلور (1953) (MAS) وهو مقياس تقرير ذاتي يتألف من 50 سؤالاً MMPI تم مقارنة أداء

الممتحنين ذوي المستويات العالية والمنخفضة من القلق، بواسطة مهام مختلفة مع فرضيات اختبار مستمدة من نظرية هوليان للتعلم (سبنس 1958). دلت الاكتشافات في هذه الدراسات أن علامات MAS العالية تنبأت بالأداء في المهام التعليمية، ولكن في الحالات التي تتطلب بعض درجات التوتر النفسي فقط (سبيلبرغ، 1966a).

وقد أظهرت أيضاً، الأبحاث المتعلقة بالقلق والتعليم أن صعوبة المهمة، والفروقات الفردية في الذكاء، والعوامل التي تؤثر على القدرات النفسية للإجابات الصحيحة والمنافسة في حالة تعليمية معينة، يجب أن تؤخذ بعين الاعتبار.

مهد كاتل وسشير (1958 - 1961) الطريق إلى تطبيقات التقنيات المتعددة التباين، وذلك لقياس شدة القلق كحالة وجدانية والفروقات الفردية في نزاع القلق كسمة شخصية (كاتل، 1961، 1963) في تحقيقات عن التغيرات المساعدة والوقت الزائد لعدد من مقاييس القلق المختلفة، برزت بشكل مترابط الحالات المستقلة نسبياً وسمات عوامل القلق (كاتل، 1966). ترافقت التغيرات النفسية مع تنشيط (تهيج) الجملة العصبية. التي تقلبت بمرور الوقت وتغيرت بشكل متباين في حالات القياس (نسبة انقطاع التنفس وضغط الدم) التي كان لديها حمولة كبيرة على عامل حالة القلق، وحمولة خفيفة على سمة عامل القلق.

إن القياسات التي ألقت بحمولة كبيرة على سمة عامل القلق المتعلق بكاتل، تضمنت تقارير ذاتية عن القلق وكانت ثابتة نسبياً بمرور الوقت. وقد ربطت علامات كاتل وسشير لمقياس القلق (1963) IPAT ومقياس سمة القلق 0.85 مع مقاييس تايلور 1953 المتعلقة بـ MAS.

قدم هذا الاكتشاف برهاناً قوياً أن MAS يقيس نزعة القلق أو سمة القلق أكثر من مجرد مستوى الباعث (الدافع). وهو بشكل مفهومي، متعلق أكثر بمستوى شدة حالة القلق في وقت ما.

اتضح أن IPAT و MAS يقيسان الفروقات الفردية للقلق والسمة الشخصية. أي انه تبين أن نزعة الإجابة إلى المواقف تكون أكثر توتراً مع تصاعد أكبر للشدة التي تساهم في تصعيد مستوى الدافع.

إن مفاهيم حالة القلق (S-Anxiety) وسمة القلق (T-Anxiety)، تشير إلى تركيبين مترابطين ولكن مختلفين تماماً منطقياً (سبيلبرغ وغراسينر، 1988). يمكن أن تُعرف حالة القلق كحالة شعورية نفسية تتألف من مشاعر شخصية للشدة والهاجس والتوتر والقلق، وهو منشط (مهيج) لآلية النظام العصبي (سبيلبرغ، 1966b، 1972). إن المقاييس الصادقة لحالة القلق تتراوح في الشدة والتقلب عبر الوقت كوظيفة لتهديد مدرك. لحالة القلق خواص لفئة من التراكيب التي وصفها اتكينسون (1964) بـ "الدوافع" ودعاها كامبيل (1963) ميولاً سلوكية مكتسبة (سبيلبرغ وديلز- غورو، 1983) تشمل خواص سمة القلق فروقات ثابتة نسبياً بين الناس للميل إلى إدراك المواقف المجهدة، كأخطار أو تهديدات أكثر أو أقل. وفي قابلية الإجابة إلى مثل هذه المواقف مع تقديرات مطابقة في سمة القلق. تقوم مقاييس حالة القلق بتقدير التكرار الذي تعرضت له حالات القلق في الماضي، واحتمال أن حالة القلق سوف تظهر في المستقبل كردة فعل على منبه مهدد (سبيلبرغ، 1983، سبيلبرغ وغراسينر، 1988).

تثبت نظرية حالة - سمة القلق أن الناس الذين لديهم مستوى عال لحالات القلق الإدراكية الاجتماعية، يقومون بتقديرها كونها أكثر تهديداً، من الناس الذين لديهم تقدير منخفض لحالات القلق (سبيلبرغ، 1972، 1972). بناء على هذا، فإن الأشخاص ذوي المستوى العالي لحالات القلق هم أكثر احتمالاً بالتعرض لارتفاعات في الشدة في حالات القلق بمثل هذه المواقف. تم تطوير بيان حالة القلق (STAI) لتقديم معايير تقارير ذاتية مختصرة نسبياً لتقدير حالة وسمة القلق في الممارسة السريرية والبحثية (سبيلبرغ وآخرون، 1970). تم التحسين والتوسع في نظرية

إشارة الخطر لفرويد (1936) ومفاهيم كاتل (1963، 1966) لحالة وسمة القلق (كاتل وسشير1958، 1961) من قبل سبيلبرغ b(1966، 1972، 1979) الذي قدم بنية المفهوم الذي أدى إلى عملية تركيب الاختبار.

تركيب بيان حالة - سمة القلق

تم تركيب معيار حالة القلق STAI لقياس التغيرات في شدة القلق كحالة شعورية في مستويات منخفضة لحالة القلق يشعر الفرد بالهدوء والأمن. إن المشاعر المتصاعدة من الشدة والهاجس والتوتر تُختبر كحالة قلق متصاعدة مع الحدود القصوى من الخوف والرعب في أعلى المستويات. وهكذا، يتضمن مفهوم حالة القلق بعداً متعلقاً بالشدة. ويتطلب مفهوم تحديد شدة السؤال الانتباه إلى حقيقة أن الأسئلة التي تقيس شدة حالة وجدانية ما، سوف تكون أكثر تأثيراً من بعض المستويات من الشدة من الأسئلة الأخرى (سبيلبرغ وآخرون، 1970)

على الرغم من أهمية تحديد الشدة للسؤال قد تم تأكيدها من قبل اناستاسي (1988)، فقد جرى تجاهل هذا المفهوم بشكل كبير، أو في أفضل الحالات، تم إدراكه بشكل هامشي فقط، في تركيب مقاييس الحالات الوجدانية والسمات الشخصية. تم إدراك أهمية تحديد الشدة في السؤال بشكل مبسط من قبل زوكيرمان (1960) الذي أدرج أسئلة فسرت المشاعر الإيجابية لقياس المستويات المنخفضة من القلق، في قائمة مراجعة التأثر الوصفي الخاصة به (STAI) (الصيغة x) لمقياس حالة القلق الذي شمل أعداداً متوازية من أسئلة وجود القلق أسئلة غياب القلق، وذلك لتسهيل قياس المجموعة الكبيرة من الشدائد النفسية (سبيلبرغ وآل، 1970) إن أسئلة STAI لغياب القلق، مثل " أنا أشعر بالراحة" تعد ذات حساسية أكبر في تقدير المستويات المنخفضة لحالة القلق، أما في أسئلة وجود القلق المترافقة مع مفاتيح كلمات مثل: الشدة أو التوتر، فإنها أكثر تأثيراً في مقياس المستويات العالية من الشدة.

كان الهدف الأساسي من تطوير STAI هو إنشاء قائمة من مجموعة واحدة من الاسئلة التي يمكن استعمالها مع التراكيب الملائمة لتقييم كم من الشدة في حالة القلق والفروقات الفردية في سمة القلق. إجابة لمقياس حالة القلق STAI، تم توجيه الممتحنين لتقدير مدى الشدة في شعورهم بالقلق (مثال: أنا أشعر بالتوتر)، "الآن في هذه اللحظة". في مقياس نسبة النقاط الأربعة التالي: كلا على الإطلاق، نوعاً ما، بشكل معتدل، بشكل كبير. وقد تطلبت تعليمات مقياس حالة القلق STAI إجابات للدلالة " كيف يشعرون بشكل عام". وذلك بسرد مقدار اختبارهم للأفكار والمشاعر المتعلقة بالقلق والأعراض الجسدية، في مقياس نسبة النقاط الأربعة التالي: "على الأغلب إطلاقاً"، "أحياناً"، "غالباً"، "على الأغلب دائماً" (سبيلبرغ، 1983).

في تقدير الثبات والترابط في صدق تركيب STAI صيغة (A) التمهيدية، تداخلت الدلالات النفسية لمفاتيح الكلمات في عدة أسئلة مع استخدامها كمقاييس لكل من حالة القلق وسمة القلق. لم يُمكن تبديل الأسئلة من التغلب على الحالة القوية أو تطبيقات السمة لمفاتيح هذه الكلمات (سبيلبرغ وآخرون، 1970) مثلاً، وجد أن "أنا أشعر بالانزعاج" كانت مقياساً عالي الحساسية لحالة القلق، كما انعكست في درجات الأسئلة العالية بشكل كبير تحت الظروف المجهدة، وعلامات أكثر انخفاضاً تحت الظروف المريحة.

علاوة على ذلك، حين تم إعطاء هذه الأسئلة مع سمات التركيب، كانت درجات هذا السؤال غير ثابتة كل الوقت، وترابطات هذه الدرجات مع أسئلة حالة القلق الأخرى كانت ضعيفة نسبياً (سبيلبرغ، 1985) على العكس من ذلك، تم ربط درجات السؤال " أنا أقلق أكثر من اللازم" بشكل كبير مع أسئلة حالة القلق الأخرى. لكنها لم تؤمن زيادة تحت الظروف التجريبية المجهدة، وفشلت في الانخفاض تحت الظروف المريحة، كما تطلب قياساً صادقاً لحالة القلق (سبيلبرغ، 1970).

بافتراض هذه الصعوبات التي تتم مواجهتها في قياس حالة القلق وسمة القلق مع الأسئلة نفسها، تم تعديل استراتيجية تركيب الاختبار لـ STAI، وتم اختيار أسئلة اختبار منفصل لقياس شدة حالة القلق كحالة شعورية، والفروقات الفردية في سمة القلق كسمة شخصية. من أجل معيار لحالة القلق (STAI) (صيغة X) تم اختيار عشرين سؤالاً مع صدق تركيب جيدة كمقاييس لحالة القلق، كما أشير إلى ترابطات كبيرة مع الـ (MAS تايلور، 1956، ومقياس القلقIPAT كاتل وسشيرو، 1963). ومقياس القلق ويلش (1956). وقد كانت درجات هذه الأسئلة ثابتة نسبياً بمرور الوقت (سبيلبرغ وآل، 1970).

جرى الاختيار من أجل تأمين معيار لحالة القلق (STAI (صيغة X) سبيلبرغ وآخرون، (1970، العشرين سؤالاً مع صدق التركيب الأفضل كمقاييس لحالة القلق، كما وُضح بواسطة الدرجات الأعلى تحت الظروف المجهدة، والدرجات الأقل تحت الظروف المريحة. حققت خمسة أسئلة فقط، معايير الصدق لكل من حالة القلق وسمة القلق، لتسمح لهم بالاندماج في كلا المعيارين. وهكذا فإن 30 من أصل 40 سؤالاً يؤلف مقياس حالة القلق ومقياس سمة القلق STAI (صيغة X)، وقد اختلفت بشكل كبير في صدق المفهوم والجدارة لتعتبر كمقاييس فريدة لكل من حالة وسمة القلق.

على أساس الحِكَم التي تم اكتسابها عبر عقد من البحث المكثف مع الـ STAI (صيغة X)، تم إجراء تنقيح أساسي لهذا المعيار (سبيلبرغ، 1983) في تركيب ومعايرة الـ STAI المنقحة (صيغة Y) تم اختبار أكثر من 5.000 ممتحن. كان الهدف الرئيس من تنقيح الـ STAI تطوير مقاييس "أنقى" لحالة القلق وسمة القلق من أجل تقديم أساس أكثر صدقاً للتفريق بين القلق والاكتئاب. نتج عن البحث الدقيق لمحتوى صدق الـ STAI (صيغة X) مع أفضل الخواص القياسية النفسية، تعريف مفهومي لتراكيب الحالة والسمة، مما قاد إلى تركيب لأسئلة تبديلية محتملة.

في تركيب الـ STAI (صيغة Y)، تم تبديل ستة أسئلة ذات محتوى بدا كأنه أكثر ارتباطاً بالاكتئاب من القلق (مثال، "أشعر بالحزن"، "أشعر بالرغبة في البكاء"). كانت الأسئلة المستبدلة أيضاً غامضة مع خواص قياسية نفسية هامشية لطلاب المدرسة الثانوية. مثل، "أشعر بالقلق" الذي فُسر من قبل العديد من هؤلاء الطلاب ليعني "يتوق" كما في "يتوق ليرضي". أسئلة مستبدلة أخرى احتوت على مصطلحات من الواضح أن معناها تغير عبر العقد المنصرم. احتمال، كنتيجة لتناول مخدر مفرط من قبل مراهقين أو يافعين صغار (مثال، أشعر أني أتخبط: "strung").

إن اختيار المجموعة الأخيرة من الأسئلة للنسخة المنقحة STAI (صيغة Y)، ارتكز على عوامل التحليل وارتباطات إشعارات الأسئلة التي نتجت عن استبدال 30٪ من الاسئلة الأصلية STAI (صيغة X) تم وضق إجراءات تبديل الأسئلة بالتفصيل في نسخة الاختبار اليدوي STAI سبيلبرغ، 1983. إن عوامل تحليل الأسئلة للنسخة المنقحة STAI (صيغة Y) حددت عوامل وسمات قلق واضحة (سبيلبرغ، فاغ، باركر، دونهام، وويسبيري، 1980، فاغ و سبيلبرغ وأوهيرن، (1980 التي ترابطت بشكل عام مع نتائج تحليلات العوامل السابقة (للصيغة X) غودري وبوول، 1975، غودري و سبيلبرغ و فاغ، 1975) لكن حالة وسمة القلق، وعوامل غياب القلق ووجود القلق في حلول الأربعة عوامل (للصيغة Y) كانت أكثر اختلافاً من العوامل الموجودة في الدراسات السابقة (للصيغة X) (سبيلبرغ، 1983). عوامل (للصيغة Y) كانت أيضاً أكثر اختلافاً، وتميزت ببنية أفضل وأكثر، بساطة وثباتاً، من العوامل المقارنة في (الصيغة X) تم أيضاً تطوير الخواص القياسية النفسية للنسخة المنقحة STAI (صيغة Y)، ومعايير حالة القلق وسمة القلق (سبيلبرغ، 1983).

تكييفات التقاطع الثقافي لاختبار SATI:

تم تكييف الـ STAI في أكثر من خمسين لغة ولهجة. واستعملت بكثرة في أبحاث التقاطع اللغوي (سبيلبرغ، سيدمان، أوين، ومارش، 1999). تم أيضاً إظهار

الترابط الداخلي والثبات والانسجام وصدق المفهوم لتكييفات اللغة الأجنبية لـ STAI سبيلبرغ ودياز-غورو، (1982)، قد قدمت براهين مدهشة على عالمية القلق كتركيب نفسي ذي مغزى. إن صيغ الـ STAI الصينية والأسكتلندية والفرنسية والألمانية والهندية والإيطالية واليابانية والبرتغالية والروسية والإسبانية، قد تم توثيقها بدقة. ونشرت معظم هذه الصيغ تجارياً. إن تكافؤ المقاييس الروسية والإنكليزية لمعايير حالة وسمة القلق، بُرهنت بواسطة ترابطات عالية أوجدت من أجل الممتحنين الثنائيي اللغة الروسية والإنكليزية. وقد أجابوا على كلا الصيغتين من هذا المقياس (هانين، 1986).

من أجل تأسيس مترجمين ومكيفين لمعايير حالة وسمة القلق، تم استخدام ميزات قياسية فريدة للغات مختلفة. في الإسبانية مثلاً هناك صيغتان لفعل الكون "ser"و "estar" تدلان ser على خاصية دائمة ثابتة نسبياً لشخص أو موقف ما، بينما لكلمة estar الدلالة على حالة انتقالية أو مؤقتة (سبيلبرغ، غونزالز- ريجوزا، مارتينيز-اوروتيا، ل. ناتاليسيو، و د. ناتاليسيو، 1971) بشكل مماثل، تتطابق الأفعال الهندية raha hun و rahta hun بشكل متعاقب مع مفاهيم حالة انتقالية وميزة ثابتة نسبياً (سبيلبرغ، شارما، وسينغ، 1973).

إن حقيقة أن اختلاف الحالة - السمة يعد شيئاً جوهرياً بالنسبة إلى اللغات الإسبانية و الهندية كما دلت بنية علم النفس اللغوي، فإن هذه الأنظمة اللغوية المختلفة ذاتها، تؤيد الحاجة الأساسية للتفريق بين الحالات الوجدانية والسمات الشخصية. علاوة على ذلك، وكما لوحظ من قبل فإن الاختلاف المتعلق بالحالة - السمة ينعكس بوضوح أيضاً في كلمات معينة لها دلالات القلق كحالة انتقالية، كالشعور "منزعج". وبإجابات على أسئلة مثل " لدي أفكار مزعجة" التي تشير ضمناً إلى استمرار ودوام السمة (سبيلبرغ، 1988).

بالإضافة إلى البرهان القوي على دلالات السمة المتوارثة في البنية اللغوية الإسبانية والهندية ولغات أخرى، وفي مفاتيح كلمات عدة أسئلة لمعايير فردية، فمن المهم أن نقدر الأنواع المفصلة من الشدة التي تحدد حالة شعورية أو سمة شخصية (أناستاسي، 1988، سبيلبرغ وآل، 1970، سبيلبرغ، شارما، 1976). بشكل مماثل لتغيرات الخطر النفسي الذي يتم تقديره عبر القياسات، مثل معدل ضربات القلب وضغط الدم، فإن معايير التقارير الذاتية لتقدير الحالات الشعورية والسمات الشخصية، يجب أن تكون حساسة تجاه التغيرات في الشدة؛ لذا، ففي تكييف مقاييس المشاعر والشخصية للتقييم التقاطع الثقافي، فإن تعريف الكلمات -في لغات مختلفة- التي تشير إلى مستويات مختلفة من تحديد شدة السؤال، يعد مطلباً أساسياً.

إن إدراج عدد تقريبي من الأسئلة في STAI صيغة (Y)، التي تصف غياب القلق، ليس فقط لتحسين الحساسية لهذا المقياس من أجل تقدير مستويات أقل لهذا التركيب، ولكن يعمل أيضاً على جعل قياس المشاعر الإيجابية شيئاً ممكناً، مثل السعادة والثقة بالنفس، التي أتضح أنها حالات شعورية متميزة وسمات شخصية. تم إدراك هذه النقطة المنهجية بوضوح في تركيب STAI-JYZ وهو التكييف الياباني لـ STAI صيغة (Y) إن STAI-JYZ تشمل عدداً متساوياً من أسئلة الحالة والسمة مع وجود القلق وغياب القلق. وتقدم مقاييس منفصلة لحالات شعورية إيجابية وسمات شخصية ترافقت مع القلق.

يجب أن نلاحظ أيضاً أن اللغات تختلف بشكل كبير في حجم مفرداتها الوجدانية. ويمكن أن تختلف أيضاً بشكل جوهري في عدد الكلمات التي تدل على إما وجود أو غياب حالة شعورية ما أو مستوى شدتها (ويرزبيكا، 1994). علاوة على ذلك، بالمقارنة مع الإنكليزية، فإن اللغات مثل الإسبانية تحتوي على مجموعة أكبر بكثير من التعابير التي تصف فروقات دقيقة للمشاعر والمستويات المترافقة لتحديد شدة السؤال (سبيلبرغ وآل، 1970). بالإضافة إلى هذا، يمكن غالباً التعبير عن

الشدة في عاطفة ما بالشكل الأوضح بواسطة مصطلحات. إن عملية تكييف عبارات المصطلحات للاستخدام في التكييفات الإسبانية لمقاييس الغضب، تمت مناقشتها أيضاً في القسم التالي:

تقييمات التقاطع الثقافي للتجربة والتعبير والسيطرة على الغضب:

في الربع الأخير من القرن الماضي، نشط الاهتمام بقياس التجربة والتعبير والسيطرة على الغضب، وذلك عبر إثبات أن الغضب والعدائية والعدوانية يرتبط بمرض ارتفاع الضغط الدموي والشرايين الدموية (دبيروسكي، ماكدوغال، ويليامز وهاني، 1984، ويليامز، بيرفوت وشيكيل، 1985). بينما تعتبر التراكيب المتعلقة بالغضب أنها غالباً ما تكون غير مترابطة وغامضة، ترتبط الخبرة والتعبير اللذان يعبران عن الغضب، بشكل نموذجي بتعريفات الغضب والعدائية والعدوانية. من الواضح أن الغضب هو العنصر الأساسي الأكبر لهذه التراكيب المتداخلة.

ارتكازاً على المراجعة الدقيقة للمطبوعات البحثية عن الغضب والعدائية والعدوانية، تم تقديم التعاريف التالية لهذه التراكيب من قبل سبيلبرغ، جاكوبس، روسيل وكرين، (1983).

يشير الغضب عادة إلى حالة شعورية تتألف من المشاعر التي تتراوح في الشدة من الغيظ المعتدل أو الانزعاج إلى الغضب والحنق. على الرغم من أن العدوانية تشمل مشاعر غاضبة، فإن لهذا المفهوم دلالات لمجموعة معقدة من المواقف التي تثير السلوكيات العدائية نحو تدمير الأشياء أو الإضرار بالناس الآخرين. يشير مفهوم العدوانية بشكل ضمني إلى السلوك المدمر أو الانتقامي نحو الأشياء أو الأشخاص الآخرين (ص160).

تم التقصي عن المظاهر السلوكية والنفسية للغضب والعدائية والعدوانية. ولكن حتى وقتنا الحالي، تم تجاهل مشاعر الغضب في الأبحاث النفسية. ونتيجة لذلك،

فإن المعايير النفسية القياسية للغضب والعدائية والعدوانية، لا تفرق عموماً بين مشاعر الغضب وتعابير الغضب والعدائية في السلوك العدواني. فشلت معظم قياسات التراكيب المتعلقة بالغضب في أخذ اختلاف الحالة-السمة بعين الاعتبار، وحصرت تجربة وتعابير الغضب بالعوامل المحددة بمواقف سلوك الغضب. إن بنية نظرية مترابطة تدرك الفرق بين الغضب والعدائية والعدوانية كتراكيب نفسية، وتميز بين الغضب كحالة شعورية وفروقات فردية في التجربة والتعبير والسيطرة على الغضب كسمات شخصية، يعد شيئاً أساسياً من أجل توجيه التركيب وتكييف التقاطع الثقافي لمقاييس الغضب.

قياس حالة وسمة الغضب:

تم تطوير بيان التعبير عن حالة - سمة الغضب (STAXI) من قبل سبيلبرغ (1988، 1999) وزملائه لقياس التجربة والتعبير والسيطرة على الغضب (سبيلبرغ وآخرون، 1983، سبيلبرغ وآخرون، 1985، سبيلبرغ وكراسنر، سولومون، 1988، سبيلبرغ وآخرون، 1995) كانت هناك أربع مراحل في تركيب وتطوير معايير STAXI الستة والمعايير الثانوية الخمسة. في المرحلة الأولية، تم تطوير مقياس حالة-سمة الغضب (STAS) لتقدير شدة الغضب كحالة شعورية واختلافات فردية في نزعة الغضب كسمة شخصية (سبيلبرغ وآخرون، 1983) تم تعريف حالة الغضب ك"حالة وجدانية تتسم بمشاعر شخصية تتراوح في الشدة من الانزعاج المعتدل أو الغيظ الى الغضب الشديد والحنق، الذي يترافق عموماً بتوتر عضلي وتهيج آلية النظام العصبي (سبيلبرغ، 1988ص1). يقدر مقياس حالة الغضب STAS مستوى شدة حالة الغضب في وقت معين.

يشير الغضب إلى اختلافات شخصية في الميل إلى اختبار مشاعر الغضب (سبيلبرغ وآخرون، (1983 يقيم معيار سمة الغضب STAS مدى التكرار الذي اختبرته حالة الغضب. إن دراسة بنية العامل المتعلق بمعيار STAS لسمة الغضب،

قد عرّف بشكل متناغم عاملين أساسين مترابطين ولكن مستقلين نسبياً: مزاج حالة الغضب وردة فعل سمة الغضب (فورجايس، فورجايس و سبيلبرغ، 1997، سبيلبرغ، 1988) قوّمت المقاييس الثانوية لحالة-مزاج الغضب بتقدير الاختلافات الفردية في الميل إلى الإحساس بمشاعر غاضبة دون إثارة. على النقيض من ذلك، فإن المقاييس لردة فعل حالة - سمة الغضب تقيس الاختلافات الشخصية في الميل إلى الإحساس والتعبير عن الغضب في المواقف التي تحتوي على إحباطات وتقديرات سلبية، أو التعرض لمعاملة غير عادلة (سبيلبرغ، 1988).

قياس التعبير والسيطرة على الغضب:

في المرحلة الثانية من تطوير الـ STAX، حرض إدراك أهمية التمييز بين التجربة والتعبير عن الغضب، على تطوير معيار التعبير عن الغضب (AX سبيلبرغ، 1985) يقوم معيار AX بتقدير عدد المرات التي تم فيها كبت الغضب (الغضب الداخلي)، أو تم التعبير عنه بسلوك عدواني (غضب خارجي). إن تعليمات الإجابة على معيار تختلف بوضوح عن تعليمات السمة التقليدية لمعيار سمة الغضب STAS بدلاً من توجيه المتحنين للإجابة تبعاً للطريقة التي يشعرون فيها بشكل عام، فقط أُعطوا تعليمات بذكر عدد المرات التي أظهروا فيها ردة فعل، أو تصرفوا بطريقة معينة حين شعروا " بالغضب أو الحنق" (مثال، "أنا أتفوه بأشياء كريهة"، "أنا أغلي من الداخل ولكني لا أظهر ذلك"). وذلك بتصنيف أنفسهم بواسطة معيار التكرار نفسه ذي النقاط الأربعة المستخدم مع مقاييس سمة الغضب.

تم تصميم معيار AX لتقدير الاتصالات الثنائية البعد، الثنائية القطب للاختلافات الفردية بعدد المرات التي تم فيها حصر الغضب (كبته)، أو عُبر عنه باتجاه الأشخاص أو الأشياء الأخرى في المحيط. إلا أن تحليلات عامل تركيب الأسئلة لقياس هذه الأبعاد بشكل مترابط حددت واحداً أو اثنين من العوامل المستقلة التي دلت أن أسئلة الـ AX استثمرت أبعاد الغضب الداخلي والغضب الخارجي (سبيلبرغ وآخرون، 1985). إن

الارتباطات بين معيار AX الداخلي و معيار AX الخارجي التي أنشئت لقياس هذه الأبعاد الضمنية، كانت أساساً صفراً (جونسون، 1984، بولانس، 1983) قدمت المزيد من البراهين أن معيار STAXI AX الداخلي ومعيار STAXI AX الخارجي، قدمت تركيبين مستقلين ومتباينين بشكل مفهومي وتجريبي.

في المرحلة الثالثة لتطوير الـ STAXI، حرض تعريف السيطرة على الغضب كعامل مستقل، على تركيب معيار لتقدير السيطرة على المشاعر الغاضبة (سبيلبرغ، 1988) إن مضمون 3 من أصل 20 سؤال أصلي لمعيار AX مثال: أسيطر على أعصابي، أحتفظ بهدوئي، أهدأ بشكل أسرع التي أدرجت لتقدير المستويات المتوسطة من التعبير عن الغضب كمعيار ثنائي القطب وثنائي البعد، قد أدت إلى ظهور جيل جديد من أسئلة إضافية تتعلق بالسيطرة على الغضب (سبيلبرغ وآخرون، 1985). إن تحليلات أسئلة السيطرة على الغضب، مع معيار AX للغضب الداخلي والخارجي، حددت عاملاً قوياً للسيطرة على الغضب وقد كان مستقلاً نسبياً عن عوامل الغضب الداخلي والخارجي. تم اختيار أسئلة السيطرة على الغضب ذات الحمولة الأكبر على عامل السيطرة على الغضب. وبشكل أساسي، حمولة الصفر على عوامل الغضب الداخلي والخارجي لمعيار (AX-CON) للسيطرة على الغضب STAXI وذلك لتقييم الاختلافات الفردية بعدد المرات التي يعمل فيها المرء بالسيطرة على التعبير الخارجي لمشاعر الغضب.

تم التحريض على المرحلة الرابعة من تركيب STAXI بسبب الأبحاث النفسية اللغوية، التي حددت الاستعارات الإنكليزية للغضب، التي بدورها دعت إلى الحاجة إلى التمييز بين آليتين للسيطرة على التعابير المتعلقة بمشاعر الغضب (لاكوف، 1987). تم وصف النموذج الأصلي لاستعارة الغضب كسائل حار في حاوية ما، حيث كان الدم هو السائل الحار وكان الجسد هو الحاوية. تعتبر شدة الغضب حالة شعورية شبيهة بالتغيرات في درجة حرارة السائل الحار. إن الاستعارة، "يغلي من

الداخل: "boiling inside، لها دلالات لمستوى شديد من الغضب المحصور. "يصدر بخاراً: blowing of steam"، تدل على التعبير الخارجي لمشاعر الغضب " يبقي الغطاء: keeping the lid on" تدل ضمناً على السيطرة على شدة الغضب، وهذا يمنع التعبير الخارجي للسلوك العدواني. وهكذا، قدمت استعارات لاكوف للغضب آليتين مختلفتين تماماً للسيطرة على الغضب: الاحتفاظ بمشاعر غاضبة مكبوتة، من أجل منع التعبير عنها، وتقليص شدة الغضب المكبوت بالهدوء.

في معيار STAXI الأصلي، كانت محتويات كل أسئلة الـ AX-CON الثمانية، ما عدا واحداً، متعلقة بالسيطرة على الغضب الخارجي (مثال، أنا أسيطر على أعصابي). لهذا، تم تركيب عدد من الأسئلة الجديدة لتقدير السيطرة على الغضب الداخلي عبر تقليص شدة الغضب المكبوت (سبيلبرغ وآخرون، 1995سيدمان، 1995) وفسرت محتويات هذه الأسئلة محاولات الهدوء والسكون والاسترخاء حين يشعر المر ء بالغضب أو الحنق. حددت تحليلات العوامل لإجابات نماذج كثيرة من الذكور والإناث الراشدين على أسئلة السيطرة على الغضب (سبيلبرغ وآخرون، 1995) عاملين للسيطرة على الغضب لكلا الجنسين: السيطرة على الغضب الداخلي والسيطرة على الغضب الخارجي.

تركيب البيان الإسباني المتعدد الثقافات للتعبير عن حالة - سمة الغضب:

إن الإسبانية ليست اللغة الناطقة في إسبانيا فقط، ولكن أيضاً في أكثر من عشرين بلداً في أمريكا الجنوبية والوسطى وفي بلاد الكاريبي. ومن قبل أكثر من 25 مليون ناطق بالإسبانية يعيشون في الولايات المتحدة. على الرغم من أن الإسبانية هي لغة أساسية في معظم أنحاء أمريكا اللاتينية ولعدد من السكان الإسبان في الولايات المتحدة، فإن الثقافات الأصلية لهؤلاء الأشخاص، غالباً ما تؤثر بشكل عميق على الإسبانية التي يتكلمون بها، وعلى تطور صفاتهم الشخصية التي تؤثر على سلوكهم. لذا، من المهم أن ندرك الاختلاف الاجتماعي والثقافي، المعقد

بشكل استثنائي، للسكان الإسبان. وأن تلك الاختلافات اللغوية بين هذه المجموعات تتغلب على عوامل الشبه. بناءً على هذا، أثناء تكييف المقاييس الإنكليزية للمشاعر والشخصية للاستخدام في الثقافات الناطقة بالإسبانية، يجب أن يبذل جهد كبير لضمان أن مفاتيح الكلمات وتعابير المصطلحات المستخدمة لتقدير المفاهيم المتعلقة بالغضب لها، جوهرياً، المعنى نفسه في المجموعات الإسبانية الثقافية المختلفة.

تم تكييف الـ STAXI-2 التجربة والتعبير والسيطرة على الغضب لسكان متباينين ثقافياً في أمريكا اللاتينية، وللثقافات الفرعية الناطقة بالإسبانية في الولايات المتحدة (موسكوسو و سبيلبرغ a 1999). من أجل هذه الغاية، تم تصميم بيان التعبير عن حالة -سمة الغضب للثقافات المتعددة الإسبانية (STAXI-SMC) جوهرياً، لقياس الأبعاد المتعلقة بالغضب التي قدرت بواسطة النسخة المنقحة لـ STAXI-2) سبيلبرغ، 1999).

تم تركيب معايير ومعايير ثانوية لتقييم الأبعاد التالية بواسطة STAXI-SMC) حالة الغضب بواسطة المعايير الثانوية لتقييم الشعور بالغضب والشعور بالرغبة بالتعبير عن الغضب. (ب) سمة الغضب بواسطة المعايير الثانوية لقياس المزاج الغاضب وردة الفعل الغاضبة. (د) معايير السمة لقياس أربعة أبعاد خاصة بالتعبير عن الغضب والسيطرة على الغضب: الغضب الداخلي، والغضب الخارجي، والسيطرة على الغضب الداخلي والخارجي (موسكوسو وسبيلبرغ، b 1999).

تم إنشاء الترجمة الأولية لأسئلة STAXI-2 من أجل STAXI-SMC وقد تم تنقيح هذه الأسئلة من قبل 26 محلل نفسي بارز من أمريكا اللاتينية. وقد تم توجيههم إلى اقتراح تعديلات وتصحيحات انطلاقاً من تفسيرات لغوية للتجربة والتعبير والسيطرة على الغضب في بلادهم (موسكوسو و سبيلبرغ b 1999). استناداً إلى آراء هؤلاء الخبراء، تم تنقيح أسئلة الـ STAXI-SMC وتم تقديم المعيار المنقح ذي الأسئلة الـ56 إلى 257 مشترك (179 امرأة، 78رجل) في المؤتمر

الخامس عشر لمجلس علم النفس بين الأمريكيتين في سان خوان بورتوريكو. احتوى النموذج على إجابات من البلاد الكاريبية (48٪)، جنوب أمريكا (32٪)، أمريكا الوسطى (16٪)، وإسبانيا (4٪). وقد تراوحت أعمارهم بين 20 و 78 عاماً (متوسط العمر= 36 سنة). وقد أكمل كل المشاركون تدريبهم في علم النفس أو تم إدراجهم في برامج التخرج أو تحت التخرج المتعلقة بعلم النفس.

إن تحليلات العوامل لإجابات الأسئلة الـ56 الأولية STAXI-SMC، أكدت الخواص البنيوية النظرية للبيان. تطابقت العوامل الثمانية المحددة بشكل جيد جداً مع العوامل المشابهة في STAXI-2، التي تضمنت عاملين من حالة الغضب وعاملين من سمة الغضب، وأربعة تعابير متعلقة بالغضب وعوامل السيطرة عليه (موسكوسو وسبيلبرغ 1999a). في تحليل العوامل المنفصلة لأسئلة حالة الغضب، تم تحديد عاملين مختلفين لكل من الرجال والنساء: "الشعور بالغضب" و "الشعور بالرغبة في التعبير عن الغضب". إلا أن اختلافات جنس الممتحنين في قوة حمولة السؤال على هذين العاملين، طرحت أسئلة مهمة تتعلق بالكيفية التي يختلف فيها الرجال والنساء من أمريكا اللاتينية في تجربة الغضب. بالنسبة للنساء، فقد فسر عامل الشعور بالغضب 73٪ من التباين الإجمالي. فبينما فسر هذا العامل 70٪ فقط من التباين الإجمالي للرجال، ولكنه فسر 13٪ فقط للنساء.

إن تحليلات العامل لأسئلة سمة الغضب STAXI-SMC التي حددت أيضاً بشكل منفصل عوامل المزاج الغاضب وردة الفعل الغاضبة. قدمت برهاناً قوياً على أن بنية العامل لهذا المقياس كانت مشابهة لتلك التي كانت في STAXI-2 إن التحليل العاملي الـ STAXI-SMC لأسئلة التعبير عن الغضب والسيطرة عليه حددت العوامل الأربعة نفسها كما في STAXI-2. إن الأسئلة المصممة لتقييم الغضب الداخلي والغضب الخارجي، والسيطرة على الغضب الداخلي والسيطرة على الغضب الخارجي، كان لها حمولات عالية على العوامل المطابقة للتعبير عن الغضب والسيطرة عليه التي كانت متشابهة عند كلا الجنسين. إن عوامل

الارتباطات لألفا لـ STAXI-SMC للمعايير والمعايير الثانوية، وسمة الغضب، ومقاييس التعبير عن الغضب والسيطرة عليه، كانت عالية بشكل معقول. لتشير إلى أن التطابق الداخلي لهذه المعايير كان مرضياً.

باختصار، إن نتائج التحليل العاملي لنتائج الممتحنين من أمريكا اللاتينية على أسئلة STAXI-SMC حددت ثمانية عوامل كانت مشابهة تماماً للأسئلة التي وجدت في STAXI-2 قامت تحليلات عاملية منفصلة لحالة الغضب وسمة الغضب على تأكيد تطابق العاملين المرتبطين، ولكن المختلفين لحالة الغضب " الشعور بالغضب" و "الشعور بالرغبة في التعبير عن الغضب". وعاملين، سمة الغضب والمزاج الغاضب وردة الفعل الغاضبة، المترابطين بشكل كبير ولكن المتباينين بوضوح. جرى أيضاً تحليلات العوامل لأسئلة التعبير عن الغضب والسيطرة عليه. وحددت نفس العوامل الأربعة الموجودة في (STAXI-2 سبيلبرغ وآخرون، 1999). وهكذا، فإن نسبة العوامل المتعددة الأبعاد لـ STAXI-SMC لإجابات الأمريكيين اللاتينيين، كانت بشكل واضح مشابهة لبنية العامل في STAXI-2 الإنكليزية.

نقاش واستنتاجات

في تكييف مقاييس الحالات الوجدانية والسمات الشخصية، تعد التراكيب النفسية غير المتكافئة في ثقافات متعددة، المصدر الرئيس للخطأ (انظر مثلاً الى: شونغ، 2004). إن تكافؤ التقاطع الثقافي يعد مسالة جدلية في تكييف مقاييس الشخصية بسبب نقص التوافق بما يتعلق بتحديد الأبعاد الأساسية للشخصية. لذا، فإن تكافؤ التقاطع الثقافي للمفاهيم التي تحدد الأبعاد التي يتم قياسها يعد شيئاً جوهرياً. يجب بذل اهتمام خاص للتمييز بين الحالات الشعورية التي تتراوح في الشدة، والفروقات الفردية في السمات الشخصية الثابتة نسبياً عبر الوقت. في تركيب الأسئلة التي تقيس الحالات الوجدانية والسمات الشخصية، من المهم أيضاً أن نأخذ بعين الاعتبار تحديد شدة السؤال، بحيث يمكن تقدير مجموعة كاملة من الشدة في حالة وجدانية ما.

تم تسهيل التكافؤ في التقاطع الثقافي لأنه اتضح أن هذه المشاعر هي نتاج عالمي للتطور. وقد لاحظ داروين أن الخوف (القلق) والحنق (الغضب) صفات شاملة لكل من البشر والحيوانات. تعدل هذه المشاعر وتحرض ردة فعل قاتل- أو- أهرب، لتساهم، كما لاحظ كانون، بالتكييف الناجح والحفاظ على البقاء. يتراوح كل من القلق والغضب في الشدة كوظيفة لكيفية ردة فعل الأفراد على الظروف المجهدة. ويتراوح الناس في الشدة والتكرار الذي اختبروا به هذه المشاعر الأساسية.

إن الكلمات المستخدمة في اللغات المختلفة لوصف الحالات الوجدانية والسمات الشخصية، متأثرة بوضوح بالاختلافات الثقافية التي تعكس المنظور الفريد لثقافة معينة بما يتعلق بالمشاعر المترافقة مع عاطفة معينة. في تكييف التقاطع الثقافي للاختبارات النفسية، يتطلب اختيار دقيق للكلمات و/ أو المصطلحات التي لها جوهرياً المعنى نفسه في كل من لغة المصدر ولغة الهدف؛ وذلك لضمان تمثيل دقيق للتراكيب النفسية التي يتم تقييمها. في تكييف مقاييس الحالات الوجدانية والسمات العاطفية، من المهم أيضاً الأخذ بالاعتبار الاختلافات الثقافية في معنى الكلمات للأشخاص الذين يتكلمون اللغة نفسها. مثال، كلمة bicho التي تعني في كوبا حشرة، ولكن تدل على قضيب الرجل في بورتوريكو.

بشكل تقليدي، تضمنت عملية تكييف الاختبارات النفسية والتربوية الترجمة الإرجاعية لأسئلة من لغة الهدف إلى لغة المصدر. على الرغم من تأكيد الترجمة الحرفية لكل كلمة، تعطي هذه الطريقة اهتماماً ضئيلاً نسبياً للتراكيب التي يتم قياسها. هناك عيبان أساسيان للترجمة الحرفية وهما، صعوبة إيجاد كلمات من لغة الهدف ذات معنى معادل لمفاتيح الكلمات في لغة المصدر. أما الثاني فهو ترجمة تعابير المصطلحات. من المهم أن يتم تكييف الدلالة الوجدانية للتعبير في لغة المصدر بدلاً من ترجمة المعنى الحرفي لكل كلمة. ومن المرغوب فيه بشدة تحديد مصطلحات ذات معنى متطابق في لغة المصدر ولغة الهدف.

تم الوصول الى تركيب واختيار أسئلة من أجل STAI عن طريق مفهوم فرويد (1963) للقلق كحالة شعورية غير سارة. وتم تعريف حالة وسمة القلق من قبل كاتل (1961، كاتل وسشير، 1961). في تركيب الـ STAI، كانت النية الأساسية هي استخدام الأسئلة نفسها مع تراكيب مختلفة لقياس حالة وسمة القلق. إن الأسئلة المنتقاة من أجل الـ STAI التمهيدية بترابط وصدق مضمون ممتازين عند قياسها للقلق. ولكن عدة أسئلة كانت. نسبياً، مقاييس ثابتة للاختلافات الشخصية في سمة القلق، عانت من نقص صدق المفهوم في تقييم حالة القلق، لأن درجات هذه الأسئلة لم تكن أعلى تحت الظروف المجهدة، ولا أدنى بعد الاسترخاء. بطريقة مماثلة، فإن عدة أسئلة ذات صدق مفهوم ممتازة كمقاييس لحالة القلق، كانت غير ثابتة عبر الوقت حين قُدمت مع تعليمات السمة.

بافتراض صعوبة استخدام الأسئلة نفسها لتقييم حالة القلق وسمة القلق، تم تعديل استراتيجية تركيب الـ STAI. كان العشرون سؤالاً الذين تم اختيارهم لسمة القلق STAI، ثابتين نسبياً" عبر الوقت. وقد تمتع أيضاً كل سؤال متعلق بسمة القلق بصدق مفهوم ممتازة. كما أشارت الترابطات المهمة مع مقاييس سمة القلق الأخرى مثل معايير MAS وIPAT المستعملة بشكل واسع.

اعتماداً على عقد من البحث المكثف لـ STAI، تم إجراء تنقيح رئيس في هذا البيان. وذلك من أجل التمييز بين القلق والاكتئاب. وقد تم استبدال الأسئلة ذات المحتوى الذي اعتبر أكثر ارتباطاً بالاكتئاب منه إلى القلق، مع الأسئلة التي تبين فيها خواص قياسية نفسية هامشية للأشخاص الأقل تعليماً. تم أيضاً تحسين التعادل بين الأسئلة التي تقدر وجود أو غياب القلق. إن التحليل العاملي لإجابات الأسئلة المنقحة من STAI صيغة (Y) حددت العوامل الأربعة التالية: حالة وجود القلق، حالة غياب القلق، سمة وجود القلق، سمة غياب القلق.

تم تكييف STAI بنجاح في أكثر من 50 لغة ولهجة. وقد تم تسهيل تكييف التقاطع الثقافي لهذا المقياس عبر حقيقة أن اختلاف حالة السمة أثبتت أنها جوهرية في البنية النفسية للغات مثل الإسبانية والهندية. وتنعكس أيضاً بوضوح في مفاتيح كلمات عدد من الأسئلة التي تحوي دلالات القلق كحالة انتقالية أو التي تصف القلق كسمة شخصية دائمة. إن إدراج عدد متساو من أسئلة غياب القلق وأسئلة وجود القلق في النسخة المعدلة لـ STAI (صيغةY) ساهم بتقييم مجموعة كاملة من الشدة في مقياس حالة القلق وسمة القلق. إن تطبيق الإجراءات المشروحة في هذا المقطع في تركيب وتطوير تكييف ياباني لـ STAI (صيغة Y) تم تفسيره من قبل فوكوهارا. في تكييف التقاطع اللغوي لمقاييس الغضب، من المهم أن يكون لدينا تعاريف مفهومية متكافئة في اللغات المترجم منها والمترجم إليها. تفرق بين الشعور بالغضب كحالة شعورية والاختلافات الفردية في التعبير والسيطرة على الغضب كسمات شخصية.

إن تركيب وتطوير بيان التعبير عن حالة وسمة الغضب الإسباني المتعدد الثقافات، تم الوصول إليه عبر تعاريف حالة وسمة الغضب والتعبير عن الغضب والسيطرة عليه. باعتبار أنه تم صياغة هذه المفاهيم في STAXI-2 وهي النسخة الإنكليزية الأصلية لهذا المقياس. وقد حددت تحليلات عوامل الأسئلة التركيبية لـ STAXI-SMC، ثمانية عوامل كانت مشابهة تماماً لبنية العامل في STAXI-2، وهكذا فإن التحليلات الإحصائية للردود على أسئلة STAXI-SMC أوضحت أن عناصر القلق المقدرة بواسطة هذا البيان مشابهة لعناصر الغضب المقدرة بواسطة STAXI-2. وتشير الدراسات حول STAXI-SMC بوضوح أن الغضب كتركيب نفسي يمكن التعرف عليه بشكل هادف كحالة وجدانية تتراوح في الشدة، وسمة شخصية معقدة مع عناصر أساسية يمكن أن تقاس بشكل تجريبي.

*****

المراجع

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Atkinson, J. (1964). *An introduction to motivation.* Princeton, NJ: Van Nostrand-Reinhold.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137–164). Newbury Park, CA: Sage.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1989). *Development and use of the MMPI–2 Content Scales.* Minneapolis: University of Minnesota Press.

Cabrera, C. (1998, August). Tricky translations: When speaking Spanish, what's acceptable in some countries could get you in trouble in others. *The Tampa Tribune,* Baylife Section, pp. 1–2.

Campbell, D. T. (1963). Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 6, pp. 94–172). New York: McGraw-Hill.

Cannon, W. (1963). *Bodily changes in pain, hunger, fear and rage.* New York, NY: Harper & Row.

Cattell, R. B. (1961). Theory of situational, instrument, second order, and refraction factors in personality structure research. *Psychological Bulletin, 58,* 160–174.

Cattell, R. B. (1963). Personality role, mood, and situation perception: An unifying theory of modulators. *Psychological Review, 70,* 1–18.

Cattell, R. B. (1966). Patterns of change: Measurements in relation to state-dimension, trait change, ability, and process concepts. In *Handbook of multivariate experimental psychology* (pp. 355–402). Chicago: Rand McNally.

Cattell, R. B., & Scheier, I. H. (1958). The nature of anxiety: A review of thirteen multivariate analyses comprising 814 variables. *Psychological Reports, 4,* 351.

Cattell, R. B., & Scheier, I. H. (1960). Stimuli related to stress, neuroticism, excitation, and anxiety response patterns. *Journal of Abnormal and Social Psychology, 60,* 195–204.

Cattell, R. B., & Scheier, I. H. (1961). *The meaning and measurement of neuroticism and anxiety.* New York: Ronald Press.

Cattell, R. B. & Scheier, I. H. (1963). *Handbook for the IPAT anxiety scale* (2nd ed.). Champaign, IL: Institute for Personality and Ability Testing.

Cheung, F. M. (2004). Use of western and indigenously-developed personality tests in Asia. *Applied Psychology: An International Review, 53*(2), 173–191.

Cohen, R. J., Swerdlik, M. E., & Smith, D. K. (1992). *Psychological testing and assessment: An introduction to tests and measurements* (2nd ed.). Columbus, OH: Mayfield.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13, 20–22.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
Darwin, C. (1965). *Expression of emotions in man and animals*. Chicago: University of Chicago Press. (Original work published 1872)
Dembroski, T. M., MacDougall, J. M., Williams, R. B., & Haney, T. L. (1984). Components of type A, hostility, and anger-in: Relationship to angiographic findings. *Psychosomatic Medicine, 47*, 219–233.
Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440.
Dimberg, U. (1994). Facial reactions: "Immediate" emotional reactions. *Psychophysiology, 31*, S40.
Dimberg, U. (1998). Fear of snakes and facial reactions: A case of rapid emotional responding. *Scandinavian Journal of Psychology, 39*, 75–80.
Ekman, P. (1973). Cross-cultural studies of facial expressions. In P. Ekman (Ed.), *Darwin and facial expression* (pp. 169–222). New York: Academic Press.
Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. San Diego: EdITS Publishers.
Forgays, D. G., Forgays, D. K., & Spielberger, C. D. (1997). Factor structure of the State-Trait Anger Expression Inventory for young adults. *Journal of Personality Assessment, 69*, 497–507.
Freud, S. (1924). *Collected papers* (Vol. 1). London: Hogarth.
Freud, S. (1936). *The problem of anxiety*. New York: Norton.
Gaudry, E., & Poole, C. (1975). A further validation of the state–trait distinction in anxiety research. *Australian Journal of Psychology, 27*, 119.
Gaudry, E., Spielberger, C. D., & Vagg, P. R. (1975). Validation of the state-trait distinction in anxiety research. *Multivariate Behavior Research, 10*, 331–341.
Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.
Guthrie, G. M., & Lonner, W. J. (1986). Assessment of personality and psychopathology. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (Vol. 8, pp. 231–264). Beverly Hills, CA: Sage.
Hall, C. S., & Lindzey, G. (1970). *Theories of personality*. New York: Wiley.
Hambleton, R. K. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.
Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1–16.
Hanin, Y. L. (1986). State–trait anxiety research on sports in the USSR. In C. D. Spielberger & R. Díaz-Guerrero (Eds.), *Cross-cultural anxiety* (Vol. 3, pp. 45–64). Washington, DC: Hemisphere.
Izard, C. E. (1977). *Human emotion*. New York: Plenum.
John, O. P. (1990). The "big five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. Pervin (Ed.), *Handbook of personality theory and research* (pp. 66–100). New York: Guilford.
Johnson, E. H. (1984). *Anger and anxiety as determinants of elevated blood pressure in adolescents*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Lakoff, G. (1987). ***Women, fire, and dangerous things: What categories reveal about the mind.*** Chicago: University of Chicago Press.

LeCompte, W. A., & Oner, N. (1976). Development of the Turkish edition of the State–Trait Anxiety Inventory. In C. D. Spielberger & R. Díaz-Guerrero (Eds.), ***Cross-cultural anxiety*** (pp. 51–67). Washington, DC: Hemisphere.

Lonner, W. J. (1990). An overview of cross-cultural testing and assessment. In R. W. Brislin (Ed.), ***Applied cross-cultural psychology*** (Vol. 14, pp. 56–76). Beverly Hills, CA: Sage.

Marsella A. J., & Leong, F. T. (1995). Cross-cultural issues in personality and career assessment. ***Journal of Career Assessment, 3***(2), 202–218.

May, R. (1977). ***The meaning of anxiety*** (Rev. ed.). New York: Norton.

Moscoso, M. S., & Spielberger, C. D. (1999a). Evaluación de la experiencia, expresión y control de la cólera en Latinoamerica [Assessing the experience, expression, and control of anger in Latin America]. ***Revista Psicologia Contemporánea, 6***(1), 4–13.

Moscoso, M. S., & Spielberger, C. D. (1999b). Measuring the experience, expression, and control of anger in Latin America: The Spanish multi-cultural State–Trait Anger Expression Inventory. ***Interamerican Journal of Psychology, 33***(2), 29–48.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. R. Scherer & P. Eckman (Eds.), ***Approaches to emotions*** (pp. 197–219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollans, C. H. (1983). ***The psychometric properties and factor structures of the Anger Expression (AX) Scale.*** Unpublished master's thesis, University of South Florida, Tampa.

Rogler, L. H. (1999). Methodological sources of cultural insensitivity in mental health research. ***American Psychologist, 54***(6), 424–433.

Spence, K. W. (1958). A theory of emotionally based drive (D) and its relation to performance in simple learning situations. ***American Psychologist, 13***, 131–141.

Spielberger, C. D. (1966a). The effects of anxiety on complex learning and academic achievement. In C. D. Spielberger (Ed.), ***Anxiety and behavior*** (pp. 361–398). New York: Academic Press.

Spielberger, C. D. (1966b). Theory and research on anxiety. In C. D. Spielberger (Ed.), ***Anxiety and behavior*** (pp. 3–20). New York: Academic Press.

Spielberger, C. D. (1972a). Anxiety as an emotional state. In C. D. Spielberger (Ed.), ***Anxiety: Current trends in theory and research*** (Vol. 2, pp. 23–49). New York: Academic Press.

Spielberger, C. D. (1979). ***Understanding stress and anxiety.*** London: Harper & Row.

Spielberger, C. D. (1983). ***Manual for the State–Trait Anxiety Inventory*** (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D. (1985). Assessment of trait and state anxiety: Conceptual and methodological issues. ***Southern Psychologist, 2***, 6–16.

Spielberger, C. D. (1988). ***State–Trait Anger Expression Inventory Manual.*** Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D. (1999). ***State–Trait Anger Expression Inventory-2.*** Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D., & Díaz-Guerrero, R. (1983). Cross-cultural anxiety: An overview. In C. D. Spielberger & R. Díaz-Guerrero (Eds.), *Cross-cultural anxiety* (Vol. 2, pp. 3–11). New York: Hemisphere/McGraw-Hill International.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L., & Natalicio, D. (1971). Development of the Spanish edition of the State–Trait Anxiety Inventory. *Interamerican Journal of Psychology, 5,* 3–4.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *STAI Manual for the State–Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Jacobs, G. A., Russell, S. F., & Crane, R. S. (1983). Assessment of anger: The State-Trait Anger Scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 159–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spielberger, C. D., Johnson, E. H., Russell, S. F., Crane, R. J., Jacobs, G. A., & Worden, T. J. (1985). The experience and expression of anger: construction and validation of an anger expression scale. In M. A. Chesney & R. H. Rosenman (Eds.), *Anger and hostility in cardiovascular and behavioral disorders* (pp. 5–30). New York: McGraw-Hill/Hemisphere.

Spielberger, C. D., & Krasner, S. S. (1988). The assessment of trait and state anxiety. In G. Burrows, R. Noyes, & M. Roth (Eds.), *Handbook of anxiety* (Vol. 2, pp. 31–51). Amsterdam: Elsevier Science.

Spielberger, C. D., Krasner, S. S., & Solomon, E. P. (1988). The experience, expression and control of anger. In M. P. Janisse (Ed.), *Health psychology: Individual differences and stress* (pp. 89–108). New York: Springer Verlag.

Spielberger, C. D., Reheiser, E. C., & Sydeman, S. J. (1995). Measuring the experience, expression, and control of anger. In H. Kassinove (Ed.), *Anger disorders: Definitions, diagnosis, and treatment* (pp. 49–67). Washington, DC: Taylor & Francis.

Spielberger, C. D., & Sharma, S. (1976). Cross-cultural measurement of anxiety. In C. D. Spielberger & R. Díaz-Guerrero (Eds.), *Cross-cultural research on anxiety* (pp. 13–25). Washington, DC: Hemisphere/Wiley.

Spielberger, C. D., & Sharma, S., & Singh, M. (1973). Development of the Hindi edition of the State–Trait Anxiety Inventory. *Indian Journal of Psychology, 48,* 11–20.

Spielberger, C. D., Sydeman, S. J., Owen, A. E., & Marsh, B. J. (1999). Measuring anxiety and anger with the State–Trait Inventory (STAI) and the State–Trait Anger Expression Inventory (STAXI). In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (pp. 993–1021). Mahwah: Lawrence Erlbaum Associates.

Spielberger, C. D., Vagg, P. R., Barker, L. R., Donham, G. W., & Westberry, L. G. (1980). The factor structure of the State–Trait Anxiety Inventory. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 7, pp. 95–109). Washington, DC: Hemisphere.

Sydeman, S. J. (1995). *The control of suppressed anger.* Unpublished master's thesis, University of South Florida, Tampa.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology, 48,* 285.

Taylor, J. A. (1956). Drive theory and manifest anxiety. *Psychological Bulletin, 53,* 303–320.

Tomkins, S. S. (1962). *Affect, imagery, and consciousness. The positive affects.* New York: Springer-Verlag.

Vagg, P. R., Spielberger, C. D., & O'Hearn, T. P., Jr. (1980). Is the State–Trait Anxiety Inventory multidimensional? *Personality and Individual Differences, 1,* 207–214.

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

Welsh, G. S. (1956). Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine* (pp. 264–281). Minneapolis: University of Minnesota Press.

Wierzbicka, A. (1994). Emotion, language, and cultural scripts. In S. E. Kitayama & H. R. M. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 133–195). Washington, DC: American Psychological Association.

Williams, R. B., Barefoot, J. C., & Shekelle, R. B. (1985). The health consequences of hostility. In M. A. Chesney & R. A. Rosenman (Eds.), *Anger and hostility in cardiovascular and behavioral disorders* (pp. 173–185). New York: Hemisphere/McGraw-Hill.

Zuckerman, M. (1960). The development of Affective Adjective Check List for the measurement of anxiety. *Journal of Consulting Psychology, 24,* 457–462.

منتدى سور الأزبكية

WWW.BOOKS4ALL.NET

https://twitter.com/SourAlAzbakya

https://www.facebook.com/books4all.net